UNIVERSIDAD CARLOS III DE MADRID

TESIS DOCTORAL

Robust estimation and outlier detection in linear models for grouped data

Autor:
Betsabé Pérez Garrido

Director/es:
Dr. Daniel Peña Sánchez de Rivera
Dr. Isabel Molina Peralta

DEPARTAMENTO DE ESTADÍSTICA

Getafe, Diciembre 2011

TESIS DOCTORAL

Robust estimation and outlier detection in linear models for grouped data

Autor: Betsabé Pérez Garrido

Director/es: Dr. Daniel Peña Sánchez de Rivera
Dr. Isabel Molina Peralta

Firma del Tribunal Calificador:

		Firma
Presidente:	Dr. Juan Romo Urroz	
Vocal:	Dr. Ralf Münnich	
Vocal:	Dr. María Dolores Ugarte Martínez	
Vocal:	Dr. Domingo Carlos Morales González	
Secretario:	Dr. María Luz Durbán Reguera	

Calificación:

Leganés/Getafe, de de

To Keán

Acknowledgements

This dissertation could not have been possible without the support of many people whom I would like to thank. First I would like to thank the great support received from my thesis advisors, Dr. Daniel Peña Sanchez de Rivera and Dr. Isabel Molina Peralta, for their support, patience and motivation over the last years. I am very grateful to Dr. Daniel Peña for his guidance throughout the development of the thesis. Special thanks go to Dr. Isabel Molina, her knowledge and experience have helped me to understand many problems presented in this dissertation.

Thanks to Dr. Roland Friend with whom I worked during June and July of 2009 in Dortmund, Germany. My colleagues Alba, Maye, Ester, Ale, Santi, Jose, etc. My old friends Lili, Lydia, Romi, Monica, Azucena, Cris and Ivonne. The Comunidad de Madrid by the research grant during the period 2006-2010. The research projects: CAM CCG06-UC3M/HUM-0866 and MEC SEJ2007-64500.

Special thanks to my parents Ma. Dolores and Gregorio for their unconditional support and for give me the best of themselves. My sister Susana, her husband Gregory and their little girls Elena and Vicky. My brother Moises, his wife Maria and their son Oscar, Katy and Attila. Finally and the most important to my beautiful son Keán and my husband Szabolcs.

Abstract

Statistical models are, implicitly or explicitly, based on certain number of assumptions. Failure of any of these assumptions can be due to the existence of atypical observations in the data that do not follow the model under consideration. In practice, the problem of outlying observations is quite common; therefore it is rather relevant to use estimation methods that appropriately treat them.

The literature provides two main alternative approaches to handle this problem. The first one consists of applying robust methods that aim to reduce the impact of outlying observations on the estimation of model parameters. The second approach attempts to use diagnostic methods that identify outlying observations before fitting the model, eliminate them and then employ a non-robust method for model estimation to the remaining clean data.

This dissertation treats the problems of robust estimation and outlier detection when data have a grouped structure and most of the data satisfy one of the following models: a linear regression model with fixed group effects or a linear regression model with random group effects.

Chapter 1 provides an introduction to the topics addressed in the dissertation, including some background information and motivation. Chapter 2 describes basic robust methods and diagnostic measures for linear regression models.

Chapter 3 introduces the linear model with fixed group effects. To reduce the impact of outlying observations, we develop an extension of the method of Peña and Yohai [34], which is based on the projection of the observations over several directions called principal sensitivity components. Outlying observations appear with extreme coordinates in these directions. Based on these coordinates, a subset of observations is chosen and an estimator based on minimizing a robust scale of the residuals (similarly to S estimators) is obtained. The new extension is called groupwise principal sensitivity components (GPSC). Our extension is compared with other proposals discussed in the literature, namely the RDL1 method proposed by Hubert and Rosseeuw [19] and the M-S estimators elaborated by Maronna and Yohai [30]. We compare these methods through different simulation scenarios and under different types of contamination. Our simulation results show that the GPSC method is able to detect a high percentage of outlying observations and a limited number of false outliers (swamping effect). It is also apt to detect outlying observations in the space of explanatory variables (called high leverage points), including the case of masked outlying observations (masking effect).

Chapter 4 introduces the linear model with random group effects, together with some diagnostic measures proposed in the literature, which are based on the assumption that the variance components are known (meaning no being estimated). In practice, variance components are not known and must be estimated from the data. Through some examples we show that the use of non-robust methods for estimating variance components can provide a wrong picture concerning the validation of model assumptions.

Chapter 5 considers a linear model with random effects for the groups. Under

this model, a robust procedure is proposed for estimation of model parameters (variance components and regression coefficients), and also for the prediction of the random effects. Variance components are estimated by a robustification of Henderson method III (Searle et al., [47]). The following benefits can be discerned related to the procedure: explicit expressions for the robust estimators are provided, avoiding iterative methods and the need for good starting values; no need for any assumption regarding the shape of the distribution of the response variable apart from the existence of first and second order moments; it is computationally low demanding; finally, the estimation procedure is simply based on the fitting of two simpler linear regression models. As a result, we propose a two-step procedure. In the first step, variance components are estimated using the robustified Henderson method III. In the second step, the fixed regression parameters are estimated and the random effects are predicted in a similar way as in Sinha and Rao [49]. This robust procedure is applied to small area estimation, in which the target is to estimate the population means of the areas. Alternative robust small area estimators are given for these means, based on the robust fitting procedure mentioned before. Chapter 6 provides an extension of the robustified Henderson method III in general linear mixed models.

Resumen

Los modelos estadísticos se basan implícita o explícitamente en un cierto número de supuestos. El incumplimiento de alguno de estos supuestos puede deberse a la existencia de observaciones atípicas en los datos que no sigan el modelo considerado. Las observaciones atípicas pueden afectar seriamente las estimaciones de los parámetros del modelo, determinando el ajuste y las predicciones. En la práctica, el problema de las observaciones atípicas es común, por tanto es importante utilizar métodos de estimación que no se vean excesivamente afectados por ellas.

En la literatura existen dos enfoques alternativos para abordar este problema. El primero consiste en el uso de métodos robustos, los cuales reduzcan el impacto de las observaciones atípicas sobre la estimación de los parámetros del modelo. El segundo consiste en el uso de métodos de diagnóstico que nos permitan identificar las observaciones atípicas antes de realizar el ajuste, descartarlas y después emplear algún método no robusto para la estimación del modelo.

En esta disertación se presentan metodologías para reducir el impacto de las observaciones atípicas sobre la estimación de los parámetros de dos modelos utilizados para modelizar datos con estructura agrupada. El primer modelo considerado es el modelo de regresión lineal con efectos fijos de los grupos y el segundo es el modelo con efectos aleatorios de los grupos.

En el Capítulo 1 se presenta una introducción sobre la motivación para abordar cada uno de los temas de esta disertación. En el Capítulo 2 se describen métodos robustos básicos y medidas de diagnosis de los modelos de regresión lineal.

En el Capítulo 3 se introduce el modelo lineal con efectos fijos de los grupos. Para reducir el impacto de las observaciones atípicas sobre este modelo, se presenta una extensión del método propuesto por Peña y Yohai [34], el cual está basado en la proyección de las observaciones sobre direcciones llamadas componentes principales de sensibilidad. Se puede demostrar que las observaciones atípicas aparecerán como coordenadas extremas sobre estas direcciones. Por tanto, una vez descartadas, es posible seleccionar un estimador basado en la minimización de una escala robusta de los residuos (esto es, similar a un estimador S). El método propuesto es llamado groupwise principal sensitivity components (GPSC). El nuevo método se compara con otras propuestas dadas en la literatura; concretamente el método RDL_1 propuesto por Hubert y Rosseeuw [19] y los estimadores M-S propuestos por Maronna y Yohai [30]. Estos métodos se comparan bajo distintos escenarios y tipos de contaminación. Los resultados muestran que el método GPSC es capaz de detectar un alto porcentaje de observaciones atípicas así como un número reducido de falsos atípicos (efecto swamping). También es apropiado para detectar observaciones atípicas en el espacio de las variables auxiliares (también llamados puntos con alto efecto palanca) así como observaciones atípicas enmascaradas (efecto masking).

En el Capitulo 4 se introduce el modelo lineal con efectos aleatorios, así como algunas medidas de diagnosis propuestas en la literatura, las cuales se basan en el supuesto de que las componentes de la varianza son conocidas (es decir, no estimadas). En la práctica las componentes de la varianza no son conocidas y

por tanto deben estimarse apartir de los datos. A través de distintos ejemplos, mostraremos que el uso de métodos no robustos para estimar las componentes de la varianza en los métodos de diagnóstico del modelo puede llevar a conclusiones erróneas en cuanto a la validación de las hipótesis del modelo.

En el Capítulo 5 se propone un procedimiento robusto para estimar los parámetros de un modelo lineal con efectos aleatorios; concretamente, las componentes de la varianza y los coeficientes de regresión, así como para predecir los efectos aleatorios. Para estimar de forma robusta las componentes de la varianza, proponemos una robustificación de los estimadores de Henderson III. Algunas ventajas de esta propuesta son las siguientes: se proveen de expresiones explícitas para los estimadores robustos, evitando el uso de métodos iterativos. Tampoco requiere de ningún supuesto sobre la forma de la distribución de la variable respuesta a excepción de la existencia de momentos hasta segundo orden; computacionalmente es menos costoso y, finalmente, la estimación de las componentes de la varianza se reduce al ajuste de modelos de regresión más simples.

Para estimar de forma robusta todos los parámetros del modelo proponemos un procedimiento a dos etapas. En la primera etapa, se estiman de forma robusta las componentes de la varianza usando la robustificación del método de Henderson III. En la segunda etapa, se estiman los coeficientes de regresión y se predicen los efectos aleatorios de forma similar a la propuesta de Sinha y Rao [49]. Después del ajuste robusto de los parámetros del modelo, se presentará una aplicación enfocada a la estimación en áreas pequeñas en la que el objetivo es la estimación de las medias de las áreas pequeñas. Se proponen unos estimadores robustos alternativos para las medias de las áreas. En el Capítulo 6 se extiende el método de Henderson III robusto al caso de un modelo lineal mixto con más de un factor aleatorio.

Contents

Chapter 1

Introduction

Linear regression models are widely used in many fields of science such as Engineering, Economics, Sociology, Health, etc. Due to the simplicity of the idea behind the least squares (LS) method, the minimization of the sum of squared residuals, and the interpretability of the final model parameter estimates, this method is very popular among practitioners. However, it is also well known that outliers, considered here as heterogeneous observations in comparison with the majority of the data, might strongly affect these estimators. Then, robust estimators are regarded as more reliable.

Robust estimation methods include those which downweight observations with extreme residuals and those that eliminate the observations pointed out by an outlier detection procedure. In the latter approach the final estimator is typically an estimator based on a clean subset of the data. Thus, these methods preserve the simplicity and the interpretability of the LS method.

On the other hand, outlier detection is an important issue itself, because singular observations might hide possibly relevant phenomena affecting our measurements. Outliers are typically pointed out using the information contained by

scaled residuals obtained from a previous model fit. However, both the scale and the previous fit used to obtain residuals might be also affected by the outliers unless they come from an initial robust fit. Thus, outlier detection and robust fitting are very related issues.

Linear regression models have received great attention in the literature of robust and diagnostic methods. However, until now little attention has been paid to linear models for data with a grouped structure.

This dissertation studies specific linear models that are used when our data are grouped according to a categorical variable. Chapter 3 studies linear models with fixed group effects. These models are typically assumed when, given constant values of auxiliary variables, the groups have different means. The number of groups is assumed to be moderate and the number of observations within each group is large enough to allow estimation of the different group means. Chapter 4 introduces linear models with random group effects. These are used to model data in which observations belonging to the same group are correlated and this correlation is constant. There are typically many groups and the sample size within some of the groups might be small. Under these two different grouped data structures, existing robust methods either might fail or cannot be applied due to computational problems. Thus, we propose new robust methods for these two situations and compare their performance with that of other available robust proposals.

Simulation results show that our robust procedure for linear models with fixed effects presents a high mean percentage of simulations with detection of 100% of true outliers while small number of observations were wrongly regarded as outliers. Particulary, when there is only contamination in the response variable

(vertical outliers), the level of the swamping effect in our robust procedure is the lowest among the compared robust methods.

In the case of linear models with random effects, simulation results show that our robust proposal for estimating variance components presents the minimum mean squared error when outlying groups are present. Moreover, the proposed robust procedure for estimating model parameters avoids the problem with starting values and it is computationally less demanding.

Chapter 2

Linear regression model

2.1 Introduction

Consider the usual linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

or in vectorial form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the vector containing the observable responses, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the $n \times p$ design matrix of full-column rank that contains the values of p variables for the n individuals or sampling units, $\boldsymbol{\beta}$ is a p-vector of unknown parameters, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is the vector of independent unobservable errors, each with zero mean and unknown variance σ^2. The first column of the design matrix $\mathbf{X}$ is composed by ones when intercept is considered in the model. The main elements of the fitting process using the method of least squares are the following:

Parameter estimates: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, with $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Projection or Hat matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, which is symmetric and idempotent.

Fitted values: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, with $E(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$ and $var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$.

Residuals: $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y}$, with $E(\hat{\boldsymbol{\epsilon}}) = \mathbf{0}_n$ and $var(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{0}_n$ denotes a vector of zeros of size n.

High leverage points: The Hat matrix

The $n \times n$ Hat or projection matrix $\mathbf{H} = (h_{ij})$, and in particular its diagonal elements h_{ii}, $i = 1, \ldots, n$, play a crucial role in the process of model diagnose. We start describing some of its properties. This matrix is symmetric and idempotent. From these two facts, it is easy to see that the sum of the squared elements of the rows (columns) are equal to the diagonal element, that is,

$$\sum_{j=1}^{n} h_{ij}^2 = h_{ii}. \tag{2.2}$$

Moreover, its eigenvalues are either zero or one and $\text{rank}(\mathbf{H}) = \text{trace}(\mathbf{H})$. Since $\text{trace}(\mathbf{H}) = \text{trace}(\mathbf{I}_p) = p$, then

$$\sum_{i=1}^{n} h_{ii} = p.$$

Thus, the average size of the diagonal elements of the Hat matrix is p/n. When the first column of $\mathbf{X}$ is a vector of ones, it holds that $1/n \leq h_{ii} \leq 1$ for every i. This last fact, together with (2.2), imply that when there is an observation i with $h_{ii} = 1$, then the rest of elements h_{ij}, $j \neq i$, in the same row (column) are equal to zero.

Different interpretations appear in the literature for the diagonal elements h_{ii} of the Hat matrix, called usually *leverages*. The first one, which explains their name, can be deduced from the relation of the predicted value of an observation and the whole set of observations,

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}\, y_j.$$

From this relation and the properties of $\mathbf{H}$ mentioned above, if there is a point i with $h_{ii} = 1$, then $\hat{y}_i = y_i$, that is, its predicted value will coincide with its observed value, in other words, the regression line will go through y_i. This means that observations with large h_{ii} values tend to lever the regression line attracting it to themselves.

Another interpretation of the leverage effect h_{ii}, which does not consider the response values, is the discrepancy of each observation $\mathbf{x}_i$ with respect to the mean $\bar{\mathbf{x}}$. Thus, points with high leverage are isolated in the space spanned by the columns of $\mathbf{X}$. A third interpretation arises from the fact that $h_{ii} = \partial\hat{y}_i/\partial y_i$. Thus, h_{ii} is the rate of variation of the predicted value $\hat{y}_i$ under an infinitesimal change in y_i, which measures somehow the influence of the response value y_i on its predicted value $\hat{y}_i$.

Hoaglin and Welsch [18] suggested a reasonable rule of thumb for considering a point as high-leverage, and this rule is when $h_{ii} > 2p/n$. Thus, high-leverage points are determined by looking at the diagonal elements of $\mathbf{H}$ and paying particular attention to any $\mathbf{x}_i$ point for which $h_{ii} > 2p/n$.

2.2 Outlier detection

Residuals describe the deviation of the observed data from the fit. Thus, an outlier in the response variable can be defined as a point $(\mathbf{x}_i^T, y_i)$ with large residual, and they can be informally detected by plotting residuals against other variables such as $\mathbf{y}$, each X_j, etc. Outlier detection should be based on standardized residuals. However, there are several ways of standardizing residuals. Below we describe the different types of standardized residuals.

It must be remarked that a high leverage point is usually associated with a small residual. This means that points that do not conform with the model and that are in an area of the $\mathbf{X}$-space with lack of points (high-leverage) will be difficult to detect using means of residuals.

a) Ordinary Residuals: The vector of ordinary residuals is $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$.

$$\hat{\boldsymbol{\epsilon}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} = (\mathbf{I}_n - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{I}_n - \mathbf{H})\boldsymbol{\epsilon} \tag{2.3}$$

This identity demonstrates clearly that the relationship between $\hat{\boldsymbol{\epsilon}}$ and $\boldsymbol{\epsilon}$ depends on $\mathbf{H}$. Thus, if the h_{ij}s are sufficiently small, then $\hat{\boldsymbol{\epsilon}}$ will serve as a reasonable substitute for $\boldsymbol{\epsilon}$, otherwise the usefulness of $\hat{\boldsymbol{\epsilon}}$ may be limited.

b) Studentized Residuals (Internal Studentization): Since $var(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I}_n - \mathbf{H})$, dividing each residual by its estimated standard deviation we obtain the standardized residuals,

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{(1 - h_{ii})}}, \quad i = 1, \ldots, n,$$

where $\hat{\sigma}^2$ is the residual mean of squares,

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\epsilon}_i^2.$$

which is an unbiased estimate of σ^2 and satisfies

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sigma^2} \sim \chi^2_{n-p}.$$

c) Studentized Residuals (External Studentization): The externally studentized residuals are defined using an estimator of σ^2 that is independent of $\hat{\epsilon}_i$. We consider as estimator of σ^2 the residual mean square error computed without the i-th case, and denoted $\sigma^2_{(i)}$. The result is the studentized residual

$$r_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}}, \quad i = 1, \ldots, n,$$

where

$$\hat{\sigma}^2_{(i)} = \frac{\Sigma_{j=1, j\neq i}^{n}(\mathbf{y}_j - \mathbf{x}_j^T\hat{\boldsymbol{\beta}}_{(i)})^2}{n-p}$$

Under normality assumptions, $\hat{\sigma}^2_{(i)}$ and $\hat{\epsilon}_i$ are independent and $r_i^* \sim t_{n-p}$.

It is possible to prefer r_i^* over r_i. The reason arises from the expression of r_i^* as function of r_i,

$$r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}},$$

which shows that r_i^* is a monotonic transformation of r_i and $r_i^{*2} \to \infty$ as $r_i^2 \to (n-p)$. Therefore, r_i^* reflects more dramatically the deviations than r_i does.

e) Predictive Residuals: Ordinary and studentized residuals (with internal studentization) are based on a fit to all the data. In contrast, the i-th predictive residual $\hat{\epsilon}_{(i)}$ is based on a fit to the data without the i-th case. Then, the i-th predicted

residual is defined by

$$\hat{\epsilon}_{(i)} = y_i - \hat{y}_{i(i)}, \quad i = 1, \ldots, n.$$

These residuals can be interpreted as prediction errors. They are used to obtain goodness of fit measures for model selection and are related with the idea of crossvalidation. They can be obtained from ordinary residuals, avoiding the n different fits, as

$$\hat{\epsilon}_{(i)} = \frac{\hat{\epsilon}_i}{1 - h_{ii}}, \quad i = 1, \ldots, n.$$

2.3 Measures of influence

This section studies the variation in the fitting results when the problem formulation is modified. For example, if a case is deleted, then results based on the reduced data set can be rather different from those based on the complete data. As Cook suggested, the study of the dependence of conclusions and inferences on various aspects of a problem formulation is known as study of influence (see e.g., Chatterjee and Hadi [7]).

Measures based on the volume of confidence ellipsoids

The following measures of influence of the i-th observation on the estimated regression coefficients are based on the change in the volume of confidence ellipsoids when i-th observation is removed from the data.

a) Andrews and Pregibon [2]. These authors argued that the deletion of a case corresponding to an outlier in Y will lead to a marked reduction in the residual sum of squares. Thus, the residual sum of squares is a diagnostic for detecting

influential cases arising due to the presence of an outlier in $\mathbf{y}$. On the other hand, the influence of a row of $\mathbf{X}$ is in part reflected by a change in the determinant of X^TX when that row is deleted. More specifically, let $\mathbf{X}^* = (\mathbf{X}, \mathbf{y})$ be the matrix of explanatory variables augmented with y. These authors suggest the relative change in the determinant,

$$AP_i = \frac{det\{\mathbf{X}_{(i)}^{*T}\mathbf{X}_{(i)}^*\}}{det\{\mathbf{X}^{*T}\mathbf{X}^*\}}$$

to analyze the influence of i-th observation.

Several remarks can be made on this measure. First, AP_i is a unitless measure. Second, $(AP_i)^{-1/2} - 1$ corresponds to the proportional change in the volume of an ellipsoid generated by $\mathbf{X}^{*T}\mathbf{X}^*$ when the i-th observation is omitted. Finally, small values of AP_i correspond to influential cases.

b) Cook and Weisberg [9]. They defined the likelihood distance as

$$LD_i = 2[L(\hat{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}}_{(i)})],$$

where $L(\hat{\boldsymbol{\beta}})$ and $L(\hat{\boldsymbol{\beta}}_{(i)})$ represent the log-likelihood evaluated at $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ respectively. The likelihood distance is related to the asymptotic confidence region

$$\left\{\boldsymbol{\beta} : 2[L(\hat{\boldsymbol{\beta}}) - L(\boldsymbol{\beta})] \leq \chi^2_{\alpha,p+1}\right\},$$

where $\chi^2_{\alpha,p+1}$ is the α critical value of the χ^2 distribution with $p+1$ degrees of freedom (p regression coefficients and the unknown variance σ^2). Due to this relation, typically LD_i is compared to χ^2_{p+1}. Observe that the definition of the likelihood distance relies on the specification of a probability distribution. For

Normal models and taking an estimator $\hat{\sigma}_u^2$ of σ_u^2, it reduces to

$$LD_i = \frac{1}{\sigma^2}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T \mathbf{X}^T \mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}).$$

c) Belsey, Kuh and Welsch [4] . These authors suggested that the influence of the i-th observation can be measured by comparing the ratio of the determinant of the estimated covariance matrix of $\hat{\boldsymbol{\beta}}_{(i)}$, when the i-th point is deleted, to the determinant of the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$, that is, to use the measure

$$CVR_i = \frac{det\{\hat{\sigma}_{(i)}^2(\mathbf{X}_{(i)}^T \mathbf{X}_{(i)})^{-1}\}}{det\{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}\}}.$$

The influence function and its sample counterparts

The basic idea of influence analysis is to introduce a small perturbation in the problem formulation, and then to monitor how the perturbation changes the outcome of the analysis. Important issues in designing methods for influence analysis are the choice of the perturbation scheme, the particular aspect of the analysis to monitor, and the method of measurement. Alternative choices for these three issues lead to different influence functions.

In the following we present some of the results concerning the influence curve. Sample versions of the influence curve provide justification for the basic tools used for finding influential cases. The influence function (IF) is defined as

$$IF_i = IF_i(\mathbf{x}_i; y_i; F; T) = \lim_{\epsilon \to 0} \frac{T[(1-\epsilon)F + \epsilon\, \delta_{\mathbf{x}_i y_i}] - T[F]}{\epsilon},$$

where $T[.]$ is a vector valued statistic based on a random sample from the probability distribution F and $\delta_{\mathbf{x}_i y_i} = 1$ at $(\mathbf{x}_i, y_i)$ and 0 otherwise. Note that IF_i measures the influence on T of adding a new observation $(\mathbf{x}_i, y_i)$ to a large sample.

Several finite sample versions of the influence curve have been suggested, three of the most promising ones are the empirical influence curve (EIC), the sample influence curve (SIC) and the sensitivity curve (SC), which are briefly described below.

a) Empirical influence curve (EIC): This curve is obtained by substituting $\hat{F}_{(i)}$ for F in the influence curve, where $\hat{F}_{(i)}$ is the empirical distribution function when the i-th observation is omitted. For linear models, taking as study statistic $\hat{\boldsymbol{\beta}}_{(i)} = T(\hat{F}_{(i)})$, we obtain

$$EIC_{(i)} = (n-1)(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{x}_i(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{(i)})$$

where $\hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{y}_{(i)}$ is the estimate of $\boldsymbol{\beta}$ obtained by removing the i-th observation. In terms of residuals, the EIC is

$$EIC_{(i)} = (n-1)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{\hat{\epsilon}_i}{(1-h_{ii})^2},$$

b) Sample influence curve (SIC): This curve is found by omitting the limit in the expression of IF_i and taking $F = \hat{F}$, $T(\hat{F}) = \hat{\boldsymbol{\beta}}$ and $\epsilon = -1/(n-1)$, obtaining

$$SIC_i = (n-1)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}).$$

In terms of residuals, the sample influence curve is

$$SIC_i = (n-1)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\frac{\hat{\epsilon}_i}{1-h_{ii}}.$$

Observe that the essential difference between EIC and SIC appears in the power of the term $(1-h_{ii})$ in the denominator.

c) Sensitivity curve (SC): This curve is obtained by setting $F = \hat{F}_{(i)}$, $T(\hat{F}_{(i)}) = \hat{\boldsymbol{\beta}}_{(i)}$ and $\epsilon = 1/n$, obtaining:

$$SC_i = n(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i).$$

Observe that SIC_i and SC_i are proportional to the distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$, given by

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \frac{\hat{\epsilon}_i}{1 - h_{ii}}.$$

Measures based on the influence function

Since the influence function IF_i for $T = \hat{\boldsymbol{\beta}}$ is a vector, it is convenient to normalize it in order to obtain a scalar measure of influence on $\hat{\boldsymbol{\beta}}$. The class of norms that are location/scale invariant is

$$D_i(M; c) = \frac{(IF_i)^T M (IF_i)}{c},$$

for an appropriate choice of matrix M and scalar c. Note that a large value of $D_i(M; c)$ indicates that the i-th observation has strong influence on the statistic T relative to M and c. There are three common choices of M and c, which lead respectively to the well-known Cook's distance, Welsch-Kuh's distance and Welsch's distance.

1) **Cook's distance:** Cook [39] proposed the use of the sample influence curve to approximate the influence function choosing the matrix $M = \mathbf{X}^T\mathbf{X}$ and the constant $c = (n-1)^2 p\, \hat{\sigma}^2$. Replacing them in $D_i(M; c)$, we obtain

$$C_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\, \hat{\sigma}^2} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}},$$

which coincides with LD_i divided by the number of explanatory variables p.

Cook also suggested to compare C_i with the quantiles of the central F distribution with p and $n-p$ degrees of freedom. C_i can be also written as

$$C_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})^T(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{p\,\hat{\sigma}^2},$$

where $\hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$ is the vector of predicted values when $\mathbf{y}_{(i)}$ is regressed on $\mathbf{X}_{(i)}$. Thus, C_i can be interpreted as the scaled euclidean distance between the two vectors of fitted values obtained by including and excluding the i-th observation.

2) **Welsch-Kuh's distance:** The impact of the i-th observation on the i-th predicted value can be measured by scaling the change in prediction at $\mathbf{x}_i^T$ when the i-th observation is omitted, that is,

$$\frac{|\hat{y}_i - \hat{y}_{i(i)}|}{\sigma\sqrt{h_{ii}}} = \frac{|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})|}{\sigma\sqrt{h_{ii}}},$$

and then using $\hat{\sigma}^2_{(i)}$ as an estimate of σ^2. Thus, the Welch-Kuh's distance is given by:

$$WK_i = \frac{|\mathbf{x}_i^T(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})|}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} = |r_i^*|\sqrt{\frac{h_{ii}}{1-h_{ii}}}.$$

3) **Welsch's distance:** Using the empirical influence curve to approximate the influence function and choosing $M = \mathbf{X}_{(i)}^T\mathbf{X}_{(i)}$ and $c = (n-1)\hat{\sigma}^2_{(i)}$, the class of norms that are location-scale invariant becomes

$$W_i^2 = D_i(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)}; (n-1)\hat{\sigma}^2_{(i)}) = (n-1)r_i^{*2}\frac{h_{ii}}{(1-h_{ii})^2}.$$

Welsch [40] suggested to use W_i as a diagnostic tool. This distance is related to the Welch-Kuh´s distance in the form

$$W_i = WK_i\sqrt{\frac{n-1}{1-h_{ii}}}$$

Observe that W_i is more sensitive than WK_i to h_{ii}. However, the fact that WK_i is easier to interpret led some authors to prefer WK_i over W_i.

2.4 Detection of groups of outliers

Some of the ideas concerning the detection of individual outliers can be extended directly to the case of multiple outlier detection. However, methods which attempt to find multiple outliers are commonly subject to phenomena called *swamping* and *masking* effects (see e.g., Simonoff and Hadi [48]).

Masking occurs when an outlier is not detected because of the presence of others; swamping when a non-outlier is wrongly considered as an outlier due to the effect of some other hidden outliers.

In this section we focus on some procedures designed to find multiple outliers in linear regression. The first class of procedures uses robust ideas to build an initial clean subset. Then, least squares estimates based on the clean subsets are combined with diagnosis ideas for outlier detection. However, for large data sets with many predictors and high leverage observations, procedures based on the clean set idea may not work well, because of the difficulty in selecting the initial subset. Other procedures are based on the eigenstructure analysis of some diagnostic matrices and are specially useful for large data sets.

1. Methods based on an initial clean set: Kianifard and Swallow [25] and [26] proposed to build a clean set of observations and compare the rest of the data with this set. If the observation closest to the clean set is not an oultlier, then increase the clean set with this observation and continue until no new observation can be incorporated into the basic set. These authors proposed to use either pre-

dictive or standarized residuals, or alternatively a measure of influence such as the Cook's distance C_i.

2. Analysis of the Influence Matrix: The matrix of changes in the predicted values is defined as:

$$\mathbf{R} = \begin{pmatrix} \hat{y}_1 - \hat{y}_{1(1)} & \hat{y}_1 - \hat{y}_{1(2)} & \dots & \hat{y}_1 - \hat{y}_{1(n-1)} & \hat{y}_1 - \hat{y}_{1(n)} \\ \hat{y}_2 - \hat{y}_{2(1)} & \hat{y}_2 - \hat{y}_{2(2)} & \dots & \hat{y}_2 - \hat{y}_{2(n-1)} & \hat{y}_2 - \hat{y}_{2(n)} \\ \dots & \dots & \dots & \dots & \dots \\ \hat{y}_{n-1} - \hat{y}_{n-1(1)} & \hat{y}_{n-1} - \hat{y}_{n-1(2)} & \dots & \hat{y}_{n-1} - \hat{y}_{n-1(n-1)} & \hat{y}_{n-1} - \hat{y}_{n-1(n)} \\ \hat{y}_n - \hat{y}_{n(1)} & \hat{y}_n - \hat{y}_{n(2)} & \dots & \hat{y}_n - \hat{y}_{n(n-1)} & \hat{y}_n - \hat{y}_{n(n)} \end{pmatrix} \tag{2.4}$$

Let us denote the columns of this matrix by $\mathbf{t}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$, $i = 1, \dots, n$. Peña and Yohai [33] presented a method to identify influential subsets by looking at the eigenvalues of an influence matrix defined as

$$\mathbf{M} = \mathbf{R}^T\mathbf{R}/p\hat{\sigma}^2$$

This matrix is defined as the uncentered covariance of a set of vectors which represent the effect on the fit of the deletion of each data point. Observe that the diagonal elements of this matrix are the Cook´s statistics. They showed that the eigenvectors of M will help to find groups of influential observations.

The Sensitivity Matrix: Now consider the rows $\mathbf{r}_i = (\hat{y}_i - \hat{y}_{i(1)}, \dots, \hat{y}_i - \hat{y}_{i(n)})$ of $\mathbf{R}$ instead of the columns. These rows indicate the sensitivity of each point, that is, how the predicted value of a given point changes when we use as sample the n sets of $n-1$ data built by deleting each point of the sample in turn. In this way, we analyze the sensitivity of a given point under a set of perturbations of

the sample. The sensitivity matrix is defined as

$$\mathbf{P} = \frac{1}{p\hat{\sigma}^2} \begin{pmatrix} \mathbf{r}_1^T\mathbf{r}_1 & \dots & \mathbf{r}_1^T\mathbf{r}_n \\ \dots & \dots & \dots \\ \mathbf{r}_1^T\mathbf{r}_n & \dots & \mathbf{r}_n^T\mathbf{r}_1 \end{pmatrix}$$

It can be shown that the sensitivity and the influence matrices have the same eigenvalues and we can obtain the eigenvectors of one matrix from the eigenvectors of the other. Peña and Yohai [34] and [33] have shown that the eigenvectors of the sensitivity matrix are more powerful tools for identifying groups of outliers than those of the influence matrix. Based on the sensitivity matrix, Peña and Yohai [34] introduced the principal sensitivity components method described in the next section.

2.4.1 The principal sensitivity components method

Peña and Yohai ([34]) proposed a fast robust procedure, called principal sensitivity components (PSC) method, for fitting a linear regression model. This method is based on outlier detection and is specially designed to detect masked high leverage outliers.

Consider the matrix of forecast changes given in (2.4) and construct the matrix:

$$\mathbf{Q} = (\mathbf{X}^T\mathbf{X})^{-1/2}(\mathbf{X}^T\mathbf{W}\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1/2}$$

where $\mathbf{W}$ is the diagonal matrix with terms $\hat{\epsilon}_i/(1-h_{ii})$.

The eigenvectors of $\mathbf{Q}$ represent the directions of maximum variability of the standarized effects

$$\gamma_i = (\mathbf{X}^T\mathbf{X})^{1/2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

To transform the effects γ_i into changes on predicted values, it is necessary to multiply γ_i by the standarized matrix $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1/2}$. Let v_i be the eigenvectors of the matrix Q. Therefore, the directions of maximum change in predicted values are obtained by premultiplying these directions $\mathbf{v}_i$ by $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1/2}$, that is

$$\mathbf{z}_i = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{v}_i$$

which represents the forecast change for each observation in the direction $\mathbf{v}_i$.

Theorem: Consider a set of regression observations $b_1 = (y_1, \mathbf{x}_1), \ldots, b_n = (y_n, \mathbf{x}_n)$ where $\mathbf{x}_i = (\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,p})^T$, $1 \leq i \leq n$, are in general position; that is, any p arbitrary points $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,p}$ are linearly independent. Suppose that we add to the sample m identical arbitrary data points $b_{n+i} = (y_{n+i}, \mathbf{x}_{n+i}) = (y*, \mathbf{x}*)$, $\mathbf{x}* = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$, $i = 1, \ldots, m$. Then, given $m < n - p + 1$ there exist k such that $||\hat{\boldsymbol{\beta}}|| > k$ and $||\mathbf{x}^*|| > k$ imply that for any set $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$, $\mathbf{v}_i = (v_{i,1}, \ldots, v_{i,n}, v_i*, \ldots, v_i*)$ of orthogonal eigenvectors of the matrix $\mathbf{RW}$, we have that

$$max_{1 \leq i \leq p} \#\{j : 1 \leq j \leq n, |v_{i,j}| \leq |v_j^*|\} > \frac{m+n}{2}$$

This theorem guarantees that high leverage outliers are expected to appear as extreme values on at least one of the principal sensitivity components $\mathbf{z}_i$.

The procedure

Here we describe a robust fitting procedure based on the PSC method. This procedure is composed of two stages, and the first stage is iterative. In the first stage, a robust estimator is obtained from a data subset that is clean of low and high leverage outliers, including groups of masked outliers. In the second stage, efficiency of the estimator is improved.

Stage 1. In this stage we find a robust estimate of $\boldsymbol{\beta}$ by an iterative procedure. In each iteration , an estimate $\hat{\boldsymbol{\beta}}^{(i)}$ is defined by

$$\hat{\boldsymbol{\beta}}^{(i)} = \text{argmin}_{\boldsymbol{\beta}\in A_i} S(\hat{\epsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\epsilon}_n(\boldsymbol{\beta})).$$

In this first iteration, the set A_1 contains $3p + 1$ elements. One of these elements is the least squares estimator. The other elements are obtained after computing the principal sensitivity components as described in Section 2.4.1. For each principal sensitivity component $\mathbf{z}_j$, $j = 1, \ldots, p$, we compute three estimates by LS as follows: the first estimate is obtained by eliminating the half of observations corresponding to the smallest coordinates of $\mathbf{z}_j$, the second eliminating the half corresponding to the largest coordinates of $\mathbf{z}_j$, and the third eliminating the half corresponding to the largest absolute values.

For the next iterations, $i > 1$ we start computing residuals $\hat{\epsilon}^{(i)} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(i-1)}$ and let $s^{(i-1)}$ be a robust scale of residuals such as the median of absolute deviations to the median (MAD). Then we delete all the observations j such that

$$|\hat{\epsilon}_j^{(i)}| \geq C_1 s^{(i-1)}.$$

Then, with the remaining observations, we compute the least squares estimator, $\hat{\boldsymbol{\beta}}_{LS}^{(i)}$, and the principal sensitivity components. The set A_i will contain $3p + 2$ elements: the new LS estimator $\hat{\boldsymbol{\beta}}_{LS}^{(i)}$, the estimate obtained in previous iteration $\hat{\boldsymbol{\beta}}^{(i-1)}$, and $3p$ estimates obtained by deleting extremes values in the principal sensitivity components similarly as in the first iteration.

The procedure ends when $\hat{\boldsymbol{\beta}}^{(i+1)} = \hat{\boldsymbol{\beta}}^{(i)}$, and the estimate that minimizes the robust scale on this stage is denoted $\hat{\boldsymbol{\beta}}_1$.

Stage 2. To gain efficiency, we define a new estimator as a one step iteration of the initial one computed in stage 1. We compute residuals $\hat{\epsilon}_j = y_j - \hat{\boldsymbol{\beta}}_1^T \mathbf{x}_j$, $1 \leq j \leq n$ and a robust scale s of the $\hat{\epsilon}_j$'s. Then we eliminate all observations j such that $|\hat{\epsilon}_j| > C_2 \cdot s$. Let n_1 be the number of observations eliminated and let $(\mathbf{y}_2, \mathbf{X}_2)$ be the sample with the $n - n_1$ remaining observations. We compute the least squares estimator, $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T \mathbf{y}_2$, and test the n_1 points previously eliminated by using the studentized out-of-sample residuals $t_j = (y_j - \hat{\boldsymbol{\beta}}_2^T \mathbf{x}_j)/\hat{s}_2 \sqrt{(1 + h_{jj})}$, where $\hat{s}_2^2 = \sum (y_j - \hat{\boldsymbol{\beta}}_2^T \mathbf{x}_j)^2/(n - n_1 - p)$ and $h_{jj} = \mathbf{x}_j^T (\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{x}_j$. Each observation in the set of $n - 1$ points is finally eliminated and considered as an outlier if $|t_j| > C_3$. With the observations that are not deleted, we compute the least squares estimator, $\hat{\boldsymbol{\beta}}$, that will be the final estimate (see Peña and Yohai [34]).

2.5 Basic robust methods

In this section we present some of the robust methods proposed in the literature for linear regression models (see Maronna and Yohai [29]). The degree of robustness of an estimate in the presence of outliers may be measured by the concept of breakdown-point which was introduced by Hampel [14]. Donoho [11] and Donoho and Huber [12] gave a finite sample version of this concept. The finite sample breakdown-point measures the maximum fraction of outliers which a given sample may contain without breaking completely the estimate (Yohai [54]).

M estimator

Huber [35] proposed a class of M-estimators that naturally generalizes the maximum likelihood estimator. Consider model (2.1) with fixed $\mathbf{X}$ where ϵ_i has density

$$\frac{1}{\sigma} f_0 \left(\frac{\epsilon}{\sigma} \right),$$

where σ is a scale parameter. For the linear model (2.1) the y_i's are independent but not identically distributed and y_i has density

$$\frac{1}{\sigma} f_0 \left(\frac{y - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right),$$

and the likelihood function for $\boldsymbol{\beta}$ assuming a fixed value of σ is

$$L(\boldsymbol{\beta}) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left(\frac{y - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma} \right),$$

Calculating the maximum likelihood estimator means maximizing $L(\boldsymbol{\beta})$, which is equivalent to finding $\hat{\boldsymbol{\beta}}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma} \right) + \log\sigma = \min, \tag{2.5}$$

where $\rho_0 = -log f_0$. We shall deal with estimates defined by (2.5). Assuming that σ is known and differentiating with respect to $\boldsymbol{\beta}$ we have the analog of the normal equations:

$$\sum_{i=1}^{n} \psi_0 \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\sigma} \right) \mathbf{x}_i = 0, \tag{2.6}$$

where $\psi_0 = \rho_0' = f_0'/f_0$. Then, the regression M-estimates of $\boldsymbol{\beta}$ are the solutions of

$$\sum_{i=1}^{n} \rho \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) = \min \tag{2.7}$$

where $\hat{\sigma}$ is an error scale estimate. Differentiating (2.7) yields the equation

$$\sum_{i=1}^{n} \psi \left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}} \right) \mathbf{x}_i = 0, \tag{2.8}$$

where $\psi = \rho'$. Solutions to (2.8) with monotone (resp. redescending) ψ are called monotone (resp. redescending) regression M-estimates. The main advantage of monotone estimates is that all solutions of (2.8) are solutions of (2.7). Furthermore, if ψ is increasing then the solution is unique.

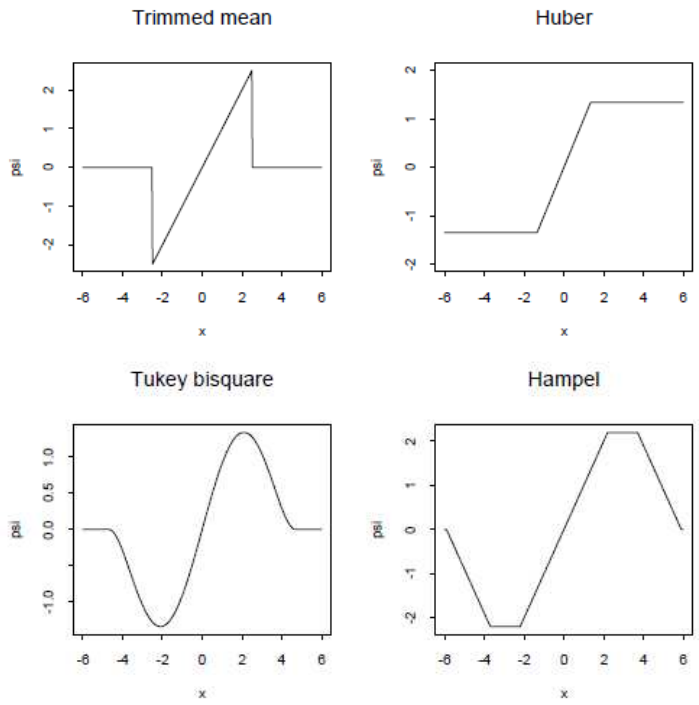


Figure 2.1: Different ψ functions for four common M-estimators.

S estimator

Yohai and Rousseeuw [45] proposed a robust estimate called S-estimator. First, consider one-dimensional estimators of scale defined by a function ρ satisfying:

a) ρ is symmetric, continuously differentiable and $\rho(0) = 0$;

b) there exist $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty]$.

For any sample $\{r_1, \ldots, r_n\}$ of real numbers, we define the scale estimate $s(r_1, \ldots, r_n)$ as the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho(r_i/s) = K \tag{2.9}$$

where K is taken to be $E_\phi[\rho(r)]$, where ϕ is the standard normal distribution function.

Now, let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be a sample of regression data with p-dimensional $\mathbf{x}_i$. For each vector $\boldsymbol{\beta}$, we obtain residuals $\hat{\epsilon}_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, of which we calculate the scale $s(\hat{\epsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\epsilon}_n(\boldsymbol{\beta}))$ by (2.9), where ρ satisfies a) and b). Then, the S-estimator $\hat{\boldsymbol{\beta}}$ is defined by

$$\text{minimize}_{\hat{\boldsymbol{\beta}}}\, s(\hat{\epsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\epsilon}_n(\boldsymbol{\beta})) \tag{2.10}$$

and the final scale estimator is

$$\hat{\sigma} = s(\hat{\epsilon}_1(\boldsymbol{\beta}), \ldots, \hat{\epsilon}_n(\boldsymbol{\beta})) \tag{2.11}$$

They decided to call $\hat{\boldsymbol{\beta}}$ S-estimator because it is derived from a scale statistic in an implicit way. S-estimators are affine equivariant, they possess a high breakdown-point and are asymptotically normal.

MM estimates

Yohai [54] proposed a new class of robust estimates called MM-estimates. The estimates have simultaneously the following properties:

a) Highly efficient when the error has a normal distribution.

b) High break-down point (concretely 50%).

MM estimates are defined by a three-stage procedure. In the first stage an initial regression estimate is computed which is consistent, robust and with high-breakdown point but not necessarily efficient. In the second stage, an M-estimate of the error scale is computed using residuals based on the initial estimate. Finally, in the third stage an M-estimate of the regression parameters based on a proper redescending psi-function is computed.

Chapter 3

Robust fitting of linear models with fixed effects

3.1 Introduction

This chapter compares several methods for outlier detection and robust estimation with grouped data, in which the majority of the data follow a linear regression model with fixed group effects. The groups might be socioeconomic population subgroups, geographical regions, strata used in the sampling scheme or, more generally, the levels of a categorical variable that is related with the outcome of interest.

Under this grouped data structure, it is possible to apply the least squares (LS) method processing the dummy variables in the same manner as the continuous ones. However the LS method is very sensitive to outliers (Hubert and Rousseeuw [19]). Another alternative is the weighted likelihood estimator (Warm [51]). Unfortunately, the method is not appropriate for grouped data. The exact fitting algorithm is computationally very expensive, whereas the algorithm based on subsampling may produce singular matrices. Moreover, the straightforward

application of a classical outlier detection procedure might lead to deletion of full groups. The application of other robust methods such as M estimation (Huber [36]) or generalized M estimation (Hampel et al. [17]) may provide very low breakdown point estimators while S estimators, which are based on minimizing a robust scale of residuals, become computationally very expensive (Maronna and Yohai [30]). Finally, the least median of squares (LMS) and the least trimmed squares (LTS) under grouped data structure might lead to singular matrices (Hubert and Rousseeuw [20]). Thus, specific methods are needed under this situation. We consider three different methods. The first is a groupwise adaptation of the principal sensitivity method of Peña and Yohai [34]. The other two are particularizations of general methods designed to fit models with continuous and categorical variables, concretely the RDL_1 method of Hubert and Rousseeuw [20] and the M-S estimator of Maronna and Yohai [30]. The three methods are compared in simulations in terms of their performance in outlier detection and their robustness.

The work is organized as follows. Section 3.2 describes the data structure and the model with fixed effects dealt with. Section 3.3 introduces the adapted principal sensitivity components method of Peña and Yohai [34] to the model with fixed group effects. Sections 3.4 and 3.5 particularize respectively the RDL_1 method of Hubert and Rousseeuw [20] and the M-S estimator of Maronna and Yohai [30] to the situation of this paper. The results of a Monte Carlo simulation study are reported in Section 3.6. An application is included in Section 3.7 and finally, some concluding remarks are given in Section 3.8.

3.2 Linear regression model with fixed group effects

Let $X = (X_1, \ldots, X_p)^T$ be a vector of continuous auxiliary variables (also called covariates) related to the study variable (also called outcome) Y. Consider that there are n sample observations of X and Y coming from D different population groups of sizes $n_1, \ldots, n_D$ with $n_d \geq 2$, $d = 1, \ldots, D$, where the overall sample size is $n = \sum_{d=1}^{D} n_d$. Let y_{dj} be the value of the study variable Y for j-th sample unit from d-th group and $\mathbf{x}_{dj} = (x_{dj1}, \ldots, x_{djp})^T$ the vector with the values of the p covariates for the same unit. In absence of outliers, we consider that sample observations follow the linear regression model

$$y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + \alpha_d + \varepsilon_{dj}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D, \tag{3.1}$$

where α_d is the effect of d-th group, assumed to be fixed, and ε_{dj} is the model error, satisfying the usual assumptions

$$\varepsilon_{dj} \sim \text{iid } N(0, \sigma^2), \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D, \tag{3.2}$$

where $\sigma^2 > 0$ is unknown. Defining the vectors $\mathbf{y}_d = (y_{d1}, \ldots, y_{dn_d})^T$ and $\boldsymbol{\varepsilon}_d = (\varepsilon_{d1}, \ldots, \varepsilon_{dn_d})^T$ and the matrix $\mathbf{X}_d = (\mathbf{x}_{d1}, \ldots, \mathbf{x}_{dn_d})^T$, the model can be expressed as

$$\mathbf{y}_d = \mathbf{X}_d \boldsymbol{\beta} + \alpha_d \mathbf{1}_{n_d} + \boldsymbol{\varepsilon}_d, \quad d = 1, \ldots, D,$$

where $\mathbf{1}_{n_d}$ denotes a vector of ones of size n_d. Here, $\boldsymbol{\varepsilon}_d \sim N(\mathbf{0}_{n_d}, \sigma^2 \mathbf{I}_{n_d})$.
Let us define the vector of group effects $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^T$. The LS estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are those satisfying

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (y_{dj} - \mathbf{x}_{dj}^T \boldsymbol{\beta} - \alpha_d)^2. \tag{3.3}$$

The estimators that satisfy the LS normal equations corresponding to (3.3) are de-

fined as follows. Consider the within group covariance matrix of the covariates,

$$\mathbf{S}_{X,d} = n_d^{-1} \sum_{j=1}^{n_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^T,$$

where $\bar{\mathbf{x}}_d = (\bar{x}_{d1}, \ldots, \bar{x}_{dp})^T$ and $\bar{x}_{dq}$ denotes the mean of the q-th auxiliary variable X_q within group d, for $q = 1, \ldots, p$. Define also the vector containing the within group covariances between each covariate and the outcome,

$$\mathbf{s}_{X,Y,d} = n_d^{-1} \sum_{j=1}^{n_d} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)(y_{dj} - \bar{y}_d),$$

where $\bar{y}_d = n_d^{-1} \sum_{j=1}^{n_d} y_{dj}$, $d = 1, \ldots, D$. Define now the combined covariance matrix $\mathbf{S}_X$ (respectively the combined covariance vector $\mathbf{s}_{XY}$) as the weighted mean of within group covariance matrices $\mathbf{S}_{X,d}$ (respectively vectors $\mathbf{s}_{X,Y,d}$) with weights proportional to the group sample sizes, i.e.,

$$\mathbf{S}_X = \sum_{d=1}^{D} \frac{n_d}{n} \mathbf{S}_{X,d}, \quad \mathbf{s}_{XY} = \sum_{d=1}^{D} \frac{n_d}{n} \mathbf{s}_{XY,d}.$$

Then, the LS estimators of $\boldsymbol{\beta}$ and α_d, $d = 1, \ldots, D$, are given by

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_X^{-1} \mathbf{s}_{XY}, \quad \hat{\alpha}_d = \bar{y}_d - \bar{\mathbf{x}}_d^T \hat{\boldsymbol{\beta}}, \quad d = 1, \ldots, D. \tag{3.4}$$

The LS estimators given in (3.4) can be alternatively obtained in two steps. Taking the mean over the units in d-th group in (3.1) we obtain $\bar{y}_d = \bar{\mathbf{x}}_d^T \boldsymbol{\beta} + \alpha_d + \bar{\varepsilon}_d$, for $d = 1, \ldots, D$, where $\bar{\varepsilon}_d = n_d^{-1} \sum_{j=1}^{n_d} \varepsilon_{dj}$. Subtracting these group means from (3.1), we obtain

$$y_{dj0} = \mathbf{x}_{dj0}^T \boldsymbol{\beta} + \varepsilon_{dj0}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D, \tag{3.5}$$

where $y_{dj0} = y_{dj} - \bar{y}_d$, $\mathbf{x}_{dj0} = \mathbf{x}_{dj} - \bar{\mathbf{x}}_d$ and $\varepsilon_{dj0} = \varepsilon_{dj} - \bar{\varepsilon}_d$, $j = 1, \ldots, n_d$, $d = 1, \ldots, D$. In the first step, we obtain the LS estimator of $\boldsymbol{\beta}$ by fitting the centered model

(3.5),

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} (y_{dj0} - \mathbf{x}_{dj0}^T \boldsymbol{\beta})^2.$$

The resulting estimator $\hat{\boldsymbol{\beta}}$ is the same as that given in (3.4). In the second step, obtain the estimator of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D)^T$ as in (3.4). The M1-S robust estimation procedure of Maronna and Yohai [30], described in Section 3.5 below, is a generalization of this two-step procedure. Predicted values are given by

$$\hat{y}_{dj} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}} + \hat{\alpha}_d, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D.$$

The vector of predicted values for group d is

$$\hat{\mathbf{y}}_d = \mathbf{X}_d \hat{\boldsymbol{\beta}} + \hat{\alpha}_d \mathbf{1}_d, \quad d = 1, \ldots, D.$$

This vector can be expressed as a linear combination of the outcome vectors for each group as $\hat{\mathbf{y}}_d = \sum_{\ell=1}^{D} \mathbf{H}_{d\ell} \mathbf{y}_\ell$, where

$$\mathbf{H}_{d\ell} = \frac{1}{n_d} \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T I(d = \ell) + (\mathbf{X}_d - \mathbf{1}_{n_d} \bar{\mathbf{x}}_d^T)(n\mathbf{S}_X)^{-1} (\mathbf{X}_\ell^T - \bar{\mathbf{x}}_\ell \mathbf{1}_{n_\ell}^T), \quad d, \ell = 1, \ldots, D.$$

Here, $I(d = \ell)$ denotes the indicator taking value 1 when $d = \ell$ and 0 otherwise. We define the hat matrix associated with d-th group as $\mathbf{H}_{dd} = (h_{jk}^d)_{j,k=1,\ldots,n_d} = \partial \hat{\mathbf{y}}_d / \partial \mathbf{y}_d^T$. The element (j, k) of this matrix measures the effect that an infinitesimal change in the outcome of k-th observation from group d has on the predicted values of j-th observation from that same group. The leverage effect of j-th observation from group d is equal to h_{jj}^d, which is here the sum of the inverse group sample size and a distance between $\mathbf{x}_{dj}$ and the group mean vector $\bar{\mathbf{x}}_d$; concretely, the leverage is given by

$$h_{jj}^d = \frac{1}{n_d} + (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d)^T (n\mathbf{S}_X)^{-1} (\mathbf{x}_{dj} - \bar{\mathbf{x}}_d), \quad d = 1, \ldots, D. \tag{3.6}$$

This indicates that observations in smaller groups have larger leverage effects than observations in larger groups, when keeping the values of the covariates the same.

The matrix $\mathbf{H}_{dd}$ is symmetric but not idempotent. If there are c_d replicates of the covariates within group d, then the elements of $\mathbf{H}_{dd}$ satisfy

$$h_{jj}^d \geq \sum_{k=1}^{n_d} h_{jk}^d h_{kj}^d = \sum_{k=1}^{n_d} (h_{jk}^d)^2 \geq c_d (h_{jj}^d)^2.$$

This, together with (3.6), implies that

$$1/n_d \leq h_{jj}^d \leq 1/c_d, \quad j = 1, \ldots, n_d, \; d = 1, \ldots, D.$$

Classical outlier detection methods are based on residuals,

$$e_{dj} = y_{dj} - \hat{y}_{dj}, \quad j = 1, \ldots, n_d, \quad d = 1, \ldots, D,$$

after an appropriate scaling. A very robust estimator of the scale, although not necessarily very efficient, is recommended to scale residuals. Still, outliers with similar values on the variables involved in the model might mask each other. Specially, groups of high leverage outliers might severely affect the final estimates, and those are exactly the ones that can not be detected by standard procedures based on residuals, due to the mentioned masking effect.

3.3 Groupwise principal sensitivity components

3.3.1 The adapted principal sensitivity components method

The PSC method cannot be directly applied to model (3.1) because it might lead to deletion of too many observations from some of the groups or even deleting a

full group. In fact, since observations in smaller groups tend to have higher leverage, these observations will be more likely deleted and then these small groups will be further reduced or even fully eliminated. Here we propose an adaptation of this method, in which each group is examined for high leverage outliers separately by computing groupwise principal sensitivity components. Thus, sensitivity vectors are defined for each group and the directions of maximum variability of these sensitivity vectors are computed for each group. Group specific principal sensitivity components are more likely to point out to outliers within the groups. Also, the procedure gives a large set of candidate estimates of the regression parameter. Minimization of a robust scale of residuals with respect to a larger set of candidate estimates makes more likely to select an estimate that is based on a initial clean subset, which in turn leads to a more robust final estimator.

We assume that at least half of the observations in each group are clean, i.e., they follow model (3.1)-(3.2). Let $\hat{y}_{dj(dk)}$ be the predicted value of y_{dj} when k-th observation from d-th group is deleted, that is

$$\hat{y}_{dj(dk)} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}_{(dk)} + \hat{\alpha}_{d(dk)}, \tag{3.7}$$

where $\hat{\boldsymbol{\beta}}_{(dk)}$ and $\hat{\alpha}_{d(dk)}$ denote respectively the LS estimates of $\boldsymbol{\beta}$ and α_d when k-th observation from d-th group is deleted (note that $\hat{\boldsymbol{\beta}}_{(dk)}$ is based on the whole sample minus the k-th observation). Similarly as in Peña and Yohai [34] but restricted to group d, for each observation y_{dj} within that group, we define the vector of changes in the predicted value when each data point from group d is eliminated, i.e.

$$(\hat{y}_{dj} - \hat{y}_{dj(d1)}, \ldots, \hat{y}_{dj} - \hat{y}_{dj(dn_d)})^T.$$

Next, we define the sensitivity matrix $\mathbf{R}_d$ for d-th group as the matrix with the

sensitivity vectors of the observations from group d in the rows, i.e.

$$\mathbf{R}_d = \begin{pmatrix} \hat{y}_{d1} - \hat{y}_{d1(d1)} & \cdots & \hat{y}_{d1} - \hat{y}_{d1(dn_d)} \\ \vdots & \ddots & \vdots \\ \hat{y}_{dn_d} - \hat{y}_{dn_d(d1)} & \cdots & \hat{y}_{dn_d} - \hat{y}_{dn_d(dn_d)} \end{pmatrix}. \tag{3.8}$$

It is easy to see that the elements of this matrix can be obtained from the leverages and the residuals of the LS fit as

$$\hat{y}_{dj} - \hat{y}_{dj(dk)} = \frac{h^d_{jk}\, e_{dk}}{1 - h^d_{kk}}, \tag{3.9}$$

avoiding to do n_d different fits. Then, the sensitivity matrix for d-th group can be expressed as $\mathbf{R}_d = \mathbf{H}_{dd}\mathbf{W}_d$, where $\mathbf{W}_d = \text{diag}_{1\leq j\leq n_d}\{e_{dj}/(1 - h^d_{jj})\}$. The matrix $\mathbf{R}_d$ has rank $p+1$, which means that the sensitivity vectors for each group lie in a subspace of dimension $p+1$. Then, similarly as in Peña and Yohai [34], the high leverage outliers within group d are expected to have extreme coordinates in at least one of the $p+1$ principal components of the sensitivity vectors. Thus, we need to obtain the eigenvectors $\{\mathbf{v}^d_q, q = 1, \ldots, p+1\}$ associated with the non null eigenvalues of matrix $\mathbf{M}_d = \mathbf{R}^T_d\mathbf{R}_d$. The maximum eigenvalue of $\mathbf{M}_d$, denoted λ^d_1, can be interpreted as the measure of the global effect of the observations of d-th group on the predicted values of the observations in that group. The eigenvector $\mathbf{v}^d_1$ associated with λ^d_1 is the direction of maximum variability of the sensitivity vectors associated with observations in d-th group. Thus, we can use the projection $\mathbf{z}^d_q = \mathbf{R}_d\mathbf{v}^d_q$ on the direction $\mathbf{v}^d_q$, $q = 1, \ldots, p+1$, to detect the high leverage points within d-th group.

As described in Section 3.3.2, the elimination of the high leverage points detected by the procedure will be followed by the detection of low leverage outliers based on residuals come from a robust estimator obtained by minimizing a robust scale

of residuals (a kind of S-estimator). Note that residuals, based on the LS fit of the model to the subset of the data which does not contain high leverage points anymore, will be suitable to detect low leverage outliers. These two consecutive steps will provide a LS estimator based on a clean subset of the data. The efficiency of this estimator will then be improved by testing for the outlyingness of each potential outlier.

Remark 3.1. Observe that $\{\mathbf{v}_q^d, q = 1, \ldots, p+1\}$ are the orthogonal directions in which the joint effect of deleting several data points in the predicted values is maximized. Also, note that since $\mathbf{W}_d$ is diagonal, the eigenvectors of $\mathbf{M}_d$ are the same as those of the within group d leverage matrix $\mathbf{H}_{dd}$, and also the same as those of $\mathbf{H}_{dd}\mathbf{H}_{dd}$.

3.3.2 The adapted robust fitting algorithm

The groupwise principal sensitivity components (GPSC) procedure described above is able to detect high leverage outliers within the groups. This procedure can be integrated in an iterative algorithm that will detect both high and low leverage outliers in each of the D groups and that will provide a final regression estimator that will be robust against those kind of outliers.

Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \alpha_1, \ldots, \alpha_D)^T$ denote the vector of regression parameters in model (3.1). The robust fitting algorithm for the model with group effects (3.1) works as follows:

Stage 1. The first iteration, $r = 1$, starts by constructing a set A_1 of candidate estimates of $\boldsymbol{\gamma}$ as follows: Obtain the sensitivity matrix $\mathbf{R}_d$ using (3.9) and compute its principal sensitivity components $\mathbf{z}_q^d$, $q = 1, \ldots, p+1$ for each group $d = 1, \ldots, D$. Now, for each component q, construct different data sets as follows. Look at

each group d and consider two different data sets from that group; in the first set include all observations from the group and in the second, delete the 50% of the observations with largest coordinates in the vector $\mathbf{d}_q^d = |\mathbf{z}_q^d - \text{median}(\mathbf{z}_q^d)|$. Combining the 2 data sets from each of the D groups we have 2^D full samples. Compute the LS estimators using each of these full samples and do the same for each of the components $q = 1, \ldots, p+1$. The LS estimates obtained from each of these full samples compose the set of candidate estimates A_1. For each candidate $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \alpha_1, \ldots, \alpha_D)^T$, obtain residuals

$$e_{dj}(\boldsymbol{\gamma}) = y_{dj} - \mathbf{x}_{dj}^T\boldsymbol{\beta} - \alpha_d, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D.$$

Then select the estimate $\boldsymbol{\gamma}^{(1)}$ satisfying

$$\boldsymbol{\gamma}^{(1)} = \underset{\boldsymbol{\gamma} \in A_1}{\text{argmin}} \quad s(e_{11}(\boldsymbol{\gamma}), \ldots, e_{Dn_D}(\boldsymbol{\gamma})), \tag{3.10}$$

where s is the normalized median absolute deviation (MAD) which is an estimator with high breakdown point. Let $\boldsymbol{\gamma}^{(r)} = ((\boldsymbol{\beta}^{(r)})^T, \alpha_1^{(r)}, \ldots, \alpha_D^{(r)})^T$ be the estimator obtained by minimizing the robust scale in iteration r. In iteration $r+1$, obtain the set of residuals associated with $\boldsymbol{\gamma}^{(r)}$,

$$e_{dj}^{(r+1)} = e_{dj}(\boldsymbol{\gamma}^{(r)}) = y_{dj} - \mathbf{x}_{dj}^T\boldsymbol{\beta}^{(r)} - \alpha_d^{(r)}, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D,$$

and let $s_d^{(r+1)} = s(e_{d1}^{(r+1)}, \ldots, e_{dn_d}^{(r+1)})^T$ be a robust scale for d-th group. For each group $d = 1, \ldots, D$, eliminate all observations with $|e_{dj}^{(r+1)}| \geq C_1 \cdot s_d^{(r+1)}$ where C_1 is a constant. With all the remaining observations from the D groups, obtain the LS estimators as in (3.4) and compute again the principal sensitivity components. Construct the set A_{r+1} with the new set of candidate estimates $\boldsymbol{\gamma}$ exactly as described before, but include in the set also the estimator obtained in previous iteration $\boldsymbol{\gamma}^{(r)}$. The iterations end when $\boldsymbol{\gamma}^{(r+1)} = \boldsymbol{\gamma}^{(r)}$ and then, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^{(r+1)} =$

$(\boldsymbol{\beta}^{*T}, \alpha_1^*, \ldots, \alpha_D^*)^T$ is called preliminary robust estimator, which is expected to be robust against possibly masked groups of high leverage points as well as low leverage outliers. This preliminary robust estimator is obtained from a possibly clean subset of data points, in which many potential outliers have been deleted. To improve the efficiency of this estimator, in Stage 2 each of these potential outliers is tested using a robust version of the t test that uses only the set of clean data points. Observations that are not rejected by this test are then returned to the sample.

Stage 2. Compute residuals from the preliminary robust estimator,

$$e_{dj}^* = e_{dj}(\boldsymbol{\gamma}^*) = y_{dj} - \mathbf{x}_{dj}^T \boldsymbol{\beta}^* - \alpha_d^*, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D,$$

and let $s_d^* = s(e_{d1}^*, \ldots, e_{dn_d}^*)$ be a robust scale for d-th group. Delete the observations with $|e_{dj}^*| > C_2 \cdot s_d^*$, where C_2 is a constant, for $d = 1, \ldots, D$. Let n^* be the total number of deleted observations. With the remaining $n - n^*$ observations, compute the LS estimators as given in (3.4) and denote them by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\alpha}_d,\ d = 1, \ldots, D$. Compute also the standard error $\tilde{\sigma}$ using the residuals of these remaining observations and the corresponding leverages $\tilde{h}_{jj}^d$. Then, test the outlyingness of each of these n^* elements by using the robust t test statistic

$$t_{dj} = \frac{y_{dj} - \mathbf{x}_{dj}^T \tilde{\boldsymbol{\beta}} - \tilde{\alpha}_d}{\tilde{\sigma}\sqrt{1 + \tilde{h}_{jj}^d}} \tag{3.11}$$

Each of the n^* observations is finally eliminated only if $|t_{dj}| > C_3$, where C_3 is a constant. The remaining observations are used to calculate the final LS estimator, denoted $\hat{\boldsymbol{\gamma}}^* = (\hat{\boldsymbol{\beta}}^{*T}, \hat{\alpha}_1^*, \ldots, \hat{\alpha}_D^*)^T$. Based on several simulation studies and a trade-off between robust and efficiency we recommend to use $C_1 = 2$ and $C_2 = C_3 = 3$.

Remark 3.2. In Stage 1, it is necessary to compute the eigenvectors of matrix M_d

of size $n_d \times n_d$. For groups d with $n_d > p + D$, this can be replaced by computing the eigenvectors of a $(p + D) \times (p + D)$ matrix. For this, define the matrices

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_D \end{pmatrix}, \quad \mathbf{Z} = \text{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_D}), \quad \mathbf{X}^* = [\mathbf{X}|\mathbf{Z}] = \begin{pmatrix} \mathbf{X}_1^* \\ \vdots \\ \mathbf{X}_D^* \end{pmatrix}. \tag{3.12}$$

It can be seen that $\mathbf{M}_d = \mathbf{\Gamma}_d \mathbf{\Gamma}_d^T$, where

$$\mathbf{\Gamma}_d = \mathbf{W}_d \mathbf{X}_d^* ((\mathbf{X}^*)^T \mathbf{X}^*)^{-1} ((\mathbf{X}_d^*)^T \mathbf{X}_d^*)^{1/2}.$$

Define now the $(p+D)\times(p+D)$ matrix $\mathbf{Q}_d = \mathbf{\Gamma}_d^T \mathbf{\Gamma}_d$. For a group d with $n_d > p+D$, it is faster to compute eigenvalues and eigenvectors of matrix $\mathbf{Q}_d$. Consider an eigenvector $\mathbf{u}_k$ of matrix $\mathbf{Q}_d = \mathbf{\Gamma}_d^T \mathbf{\Gamma}_d$ associated with eigenvalue λ_k. Then, the eigenvector of $\mathbf{M}_d = \mathbf{\Gamma}_d \mathbf{\Gamma}_d^T$ associated with the same eigenvalue λ_k is equal to $\mathbf{v}_k = \mathbf{\Gamma}_d \mathbf{u}_k$. Then, the principal sensitivity component associated with $\mathbf{v}_k$ is the projection of the rows of $\mathbf{R}_d$ on $\mathbf{v}_k$, which is equal to

$$\mathbf{z}_k = \mathbf{R}_d \mathbf{v}_k = \mathbf{R}_d \mathbf{\Gamma}_d \mathbf{u}_k = \lambda_k \mathbf{X}_d^* ((\mathbf{X}_d^*)^T \mathbf{X}_d^*)^{-1/2} \mathbf{u}_k.$$

Remark 3.3. Another way of speeding up the GPSC fitting algorithm, specially for large D, is the following. In Stage 1, after computing the $p + 1$ PSCs $\mathbf{z}_q^d$, $q = 1, \ldots, p + 1$, for each group d, instead of considering the two data sets obtained by deleting 0% and 50% of observations with largest coordinates in $\mathbf{d}_q^d$ within each group d, we can just consider the data set obtained by deleting 50% of observations with largest coordinates in $\mathbf{d}_q^d$ within each group d. This would be done for each component $q = 1, \ldots, p + 1$. Then, in the first iteration of the algorithm, the set of candidate estimates A_1 would have only $p + 2$ elements. Forcing the deletion of 50% of observations could in principle affect the efficiency

of the algorithm, but Stage 2 would then improve the estimator by returning to the sample the observations that are not really outliers.

Remark 3.4. The final estimator $\hat{\boldsymbol{\gamma}}^* = \hat{\boldsymbol{\gamma}}^*(\mathbf{X}^*, \mathbf{y})$ obtained from Stage 2 is regression and scale equivariant, that is, if we transform $\mathbf{y}$ by $\lambda\mathbf{y} + \mathbf{X}^*\boldsymbol{\delta}$, where $\lambda \in I\!R$ and $\boldsymbol{\delta} \in I\!R^{p+D}$, then

$$\hat{\gamma}^*(\mathbf{X}^*, \lambda\mathbf{y} + \mathbf{X}^*\boldsymbol{\delta}) = \lambda\hat{\gamma}^*(\mathbf{X}^*, \mathbf{y}) + \boldsymbol{\delta}.$$

It is also affine equivariant when transforming the matrix of covariates $\mathbf{X}$ by $\mathbf{XA}$, where $\mathbf{A}$ is a nonsingular $p \times p$ matrix.

3.4 RDL$_1$ method

Hubert and Rousseeuw [20] proposed the RDL$_1$ method, which consists of using a robust distance to downweight high leverage points, and then using those weights to obtain a weighted L1 regression estimator. This method works as follows:

1) First, search for high leverage points in the set

$$\mathcal{X} = \{\mathbf{x}_{dj},\ j = 1, \ldots, n_d,\ d = 1, \ldots, D\},$$

by computing the minimum volume ellipsoid (MVE) of Rousseeuw [37]. The idea is to consider all ellipsoids of approximately 50% of the observations and then select the one with smallest volume. The mean vector and the covariance matrix of that ellipsoid are considered as robust location and scatter matrix, $M(\mathcal{X})$ and $C(\mathcal{X})$ respectively, of the set of data points $\mathcal{X}$.

Then, compute the robust distances of each observation to the location as

$$RD(\mathbf{x}_{dj}) = \sqrt{(\mathbf{x}_{dj} - M(\mathcal{X}))C(\mathcal{X})^{-1}(\mathbf{x}_{dj} - M(\mathcal{X}))^T}, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D.$$

Observations with large robust distances are regarded as high leverage points. A possible disadvantage of this method is that it suffers from the swamping effect. This problem will be illustrated in the simulation study of Section 3.6.

2) Estimate the regression parameter $\boldsymbol{\gamma} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$ by a weighted L_1 regression, that is, by solving the problem

$$\min_{\boldsymbol{\gamma}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} w_{dj} |e_{dj}(\boldsymbol{\gamma})|,$$

where the weights are given by

$$w_{dj} = \min \left\{ 1, \frac{p}{RD(\mathbf{x}_{dj})^2} \right\}, \quad j = 1, \ldots, n_d,\ d = 1, \ldots, D.$$

3) Let $\hat{\boldsymbol{\gamma}}$ be the estimate obtained by the weighted L_1 regression in Step 2. Finally, following the recomendation of Maronna and Yohai [30] we compute the normalized median absolute deviation (MAD) of the nonull residuals, as

$$\hat{\sigma} = 1.4826 \cdot \text{median}\{|e_{dj}(\hat{\boldsymbol{\gamma}})|,\ j = 1, \ldots, n_d,\ d = 1, \ldots, D\} \quad \text{where} \quad e_{dj}(\hat{\boldsymbol{\gamma}}) \neq 0.$$

Under this method, an observation is classified as an outlier if its corresponding absolute standardized residual, $e_{dj}(\hat{\boldsymbol{\gamma}})/\hat{\sigma}$, exceeds 2.5.

3.5 M-S estimator

Maronna and Yohai [30] proposed an alternating M and S estimator for models that include categorical variables, where an M estimator is used for the vector of parameters of the categorical predictors and an S estimator is used for the parameters of the continuous ones. The particularization of this method to model (3.1) is defined as follows. Assume first that $\boldsymbol{\beta}$ is known. Then, obtain an M estimator of $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha}(\boldsymbol{\beta}) = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{d=1}^{D} \sum_{j=1}^{n_d} \rho(y_{dj} - \mathbf{x}_{dj}^T \boldsymbol{\beta} - \alpha_d), \tag{3.13}$$

where ρ is an even convex function. Consider the vectors of residuals

$$\mathbf{e}_d(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \mathbf{y}_d - \mathbf{x}_{dj}^T \boldsymbol{\beta} - \alpha_d \mathbf{1}_{n_d}, \quad d = 1, \ldots, D.$$

Then, the estimator of $\boldsymbol{\beta}$ is obtained by minimizing a robust scale s of the residuals obtained using the M estimator $\boldsymbol{\alpha}(\boldsymbol{\beta})$, that is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ s(\mathbf{e}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\beta})), \ldots, \mathbf{e}_D(\boldsymbol{\beta}, \boldsymbol{\alpha}(\boldsymbol{\beta}))).$$

Maronna and Yohai [30] proposed also a computationally simpler variation of the original M-S method called M1-S. This method is a generalization of the two step LS procedure described at the end of Section 3.2, based on estimating $\boldsymbol{\beta}$ after removing the effect of $\boldsymbol{\alpha}$. For this, first center the outcomes as $\mathbf{y}_{d0} = \mathbf{y}_d - t_d \mathbf{1}_{n_d}$, where t_d is an M estimator of location of the observations from d-th group $\mathbf{y}_d = (y_{d1}, \ldots, y_{dn_d})^T$, that is,

$$t_d = \underset{\alpha_d}{\operatorname{argmin}} \sum_{j=1}^{n_d} \rho(y_{dj} - \alpha_d),\ d = 1, \ldots, D.$$

Center also the rows of $\mathbf{X}_d$ as $\mathbf{X}_{d0} = \mathbf{X}_d - \mathbf{1}_{n_d} \mathbf{t}_d^T$, where $\mathbf{t}_d = (t_{d1}, \ldots, t_{dp})^T$ and t_{dq}

is an M estimator of location for the q-th column of matrix $\mathbf{X}_d$, that is,

$$t_{dq} = \underset{\alpha_d}{\text{argmin}} \sum_{j=1}^{n_d} \rho(x_{djq} - \alpha_d),\ q = 1, \ldots, p,\ d = 1, \ldots, D.$$

Finally, the estimator of $\boldsymbol{\beta}$ is obtained by fitting the centered model using an S estimator, i.e.

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ s(\mathbf{y}_{10} - \mathbf{X}_{10}\boldsymbol{\beta}, \ldots, \mathbf{y}_{D0} - \mathbf{X}_{D0}\boldsymbol{\beta}).$$

Assuming that the columns of matrices $\mathbf{X}$ and $\mathbf{Z}$ defined in (3.12) are linearly independent sets, the M1-S estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$ of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ are defined as

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}, \quad \hat{\alpha}_d = t_d - \mathbf{t}_d^T \hat{\boldsymbol{\beta}}, \quad d = 1, \ldots, D.$$

Observe that when the function ρ introduced in (3.13) is the L_1 norm $\rho(x) = |x|$, the M estimator of $\boldsymbol{\alpha}$ obtained by solving (3.13) is given by $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_D)^T$, where $\hat{\alpha}_d = \text{median}\{\mathbf{y}_d - \mathbf{X}_d\boldsymbol{\beta}\}$, $d = 1, \ldots, D$. Similarly, for the L_1 norm, $t_d = \text{median}\{\mathbf{y}_d\}$ and $t_{dq} = \text{median}\{x_{d1q}, \ldots, x_{dn_dq}\}$, for each auxiliary variable $q = 1, \ldots, p$, and for each group d, $d = 1, \ldots, D$. Although M1-S estimators are attractive due to their simplicity, they are neither regression nor affine equivariant, whereas M-S estimators are.

Maronna and Yohai [30] introduced also an estimator called M-GM for models with categorical variables. This estimator is a weighted L_1 regression estimator similar to the RDL$_1$, but in this case the weights w_{dj} are function of a measure of the outlyingness of the previously centered data points $\mathbf{x}_{dj0}$. In a simulation experiment carried out by these authors, this estimator broke down when the number of continuous predictors was greater than 3, while the M-S estimator resisted. Thus, they recommended the latter for $p > 3$.

3.6 Simulation experiment

Typically, when sample sizes grow, the effect on the final estimators of a limited number of finite outliers goes to zero. Thus, instead of studying large sample properties, it seems more convenient to study the performance of robust methods under limited sample sizes, which in turn is a much more realistic setup.

This section reports the results of a simulation experiment designed to compare the outlier detection performance and the robustness of the three procedures introduced here, namely the groupwise principal sensitivity components (GPSC), the RDL_1 and the M-S methods, under finite group sample sizes. For this, we simulated data trying to imitate a data set from the Australian Agricultural and Grazing Industries Survey (AAGIS) and used in Chambers and Tzavidis [6] and Chambers et al. [5]. This data set contains several variables measured to 1652 Australian farms. Among these variables, we find the total cash receipts of the farm business over the surveyed year (*income*), the total area of the farm (*hectares*), the area of crops grown on the farm (*crops*), the number of beef cattle on the farm (*beef*) and the number of sheep (*sheep*).

Thus, in our simulation study, data corresponding to $D = 10$ groups with a total sample size of $n = 400$ were generated. The group sample sizes were respectively $n_{2k-1} = n_{2k} = 10k + 10$, $k = 1, 2, \ldots, 5$. We generated observations corresponding to four covariates with respective distributions $X_1 \sim N(3.31, 0.68)$, $X_2 \sim N(1.74, 1.23)$, $X_3 \sim N(1.70, 1.65)$ and $X_4 \sim N(2.41, 2.61)$, where the given means and the standard deviations were taken as the sample means and standard deviations of the variables *hectares*, *crops*, *beef* and *sheep* respectively, of the AAGIS data. The true values of regression coefficients are taken as $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4) = (0.45, 0.14, 0.05, 0.005)$, obtained by fitting the fixed effects model to AAGIS data.

The fixed effects α_d, $d = 1, \ldots, 10$, were generated from a normal distribution with zero mean and standard deviation $\sigma_\alpha = 1$. The errors ε_{dj} were generated independently from a normal distribution with zero mean and standard deviation equal to $\sigma = 0.1$. Then, keeping the α_ds and the values of the covariates fixed, we carried out $L = 500$ Monte Carlo replicates. In each replicate, we generated the model responses y_{dj} from model (3.1). Then, we considered three contamination scenarios:

A. *No contamination.*

B. *Only vertical outliers*: A subset $\mathcal{D}_c \subseteq \{1, 2, \ldots, D\}$ of the groups was selected for contamination. Within these selected groups $\mathcal{D}_c$, a given percentage of the observations were contaminated as follows. For selected group $d \in \mathcal{D}_c$, half of the contaminated observations were replaced by $c_{d1} = \bar{y}_d + k\, s_{Y,d}$ and the other half to $c_{d2} = \bar{y}_d - k\, s_{Y,d}$ with $k = 5$, where $\bar{y}_d$ and $s_{Y,d}$ are respectively the mean and the standard deviation of the generated clean outcomes in d-th group. In this way, the contaminated observations are clearly outliers as compared with the clean ones.

C. *Leverage points and vertical outliers*: Again, a percentage of contamination was introduced in each selected group $d \in \mathcal{D}_c$. The contamination over the set of covariates X_q, $q = 1, 2, 3, 4$, was created marginally for each q and similarly as before, setting x_{djq} equal to $c_{d3} = \bar{x}_{dq} + k\, s_{X_q,d}$, where $\bar{x}_{dq}$ and $s_{X_q,d}$ are respectively the mean and standard deviation of the clean data of X_q in d-th group and taking $k = 5$. Finally, the responses y_{dj} corresponding to half of these observations were set to $c_{d4} = \bar{y}_d + k\, s_{Y,d}$ and the other half to $c_{d5} = \bar{y}_d - k\, s_{Y,d}$, similarly as described in scenario B.

We selected for contamination three groups of different sizes, concretely we took $\mathcal{D}_c = \{1, 5, 7\}$. Figures 3.1 and 3.2 illustrate graphically contamination scenarios B and C respectively. Figure 3.1 shows the simulated observations obtained

from one of the Monte Carlo replicates under 15% of contamination type B within selected groups. The top left plot shows an index plot of the outcomes of all the $n = 400$ generated observations. The other three plots show only the observations from each of the three contaminated groups. Observe that only vertical outliers appear.

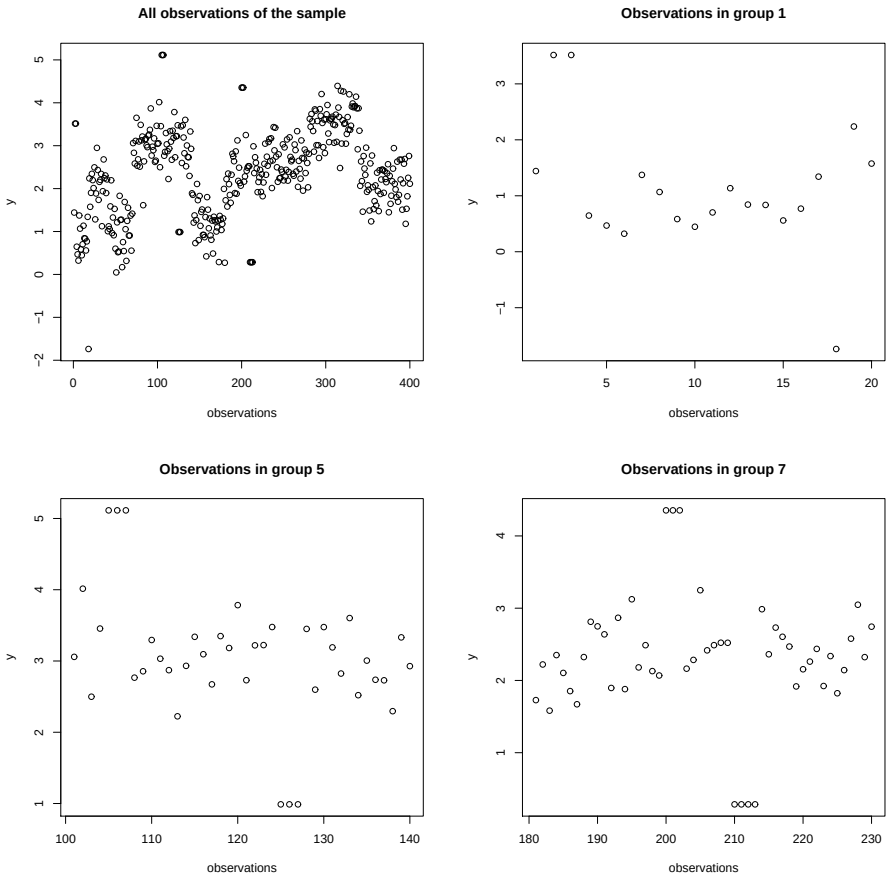


Figure 3.1: Index plots of outcomes for all observations of the sample (top left), for observations from group 1 (top right), group 5 (bottom left) and group 7 (bottom right)

Figure 3.2 shows graphically the data with 15% of contamination type C. The four plots in this figure show the outcomes of all sample observations against their values in the covariates X_q, for $q = 1, 2, 3, 4$, respectively.

Thus, for each iteration $l = 1, \ldots, L$, the three different procedures, namely GPSC, RDL_1 and M-S, where applied to the simulated data. Four main performance criteria were used to compare the results of these estimators, the first two are used

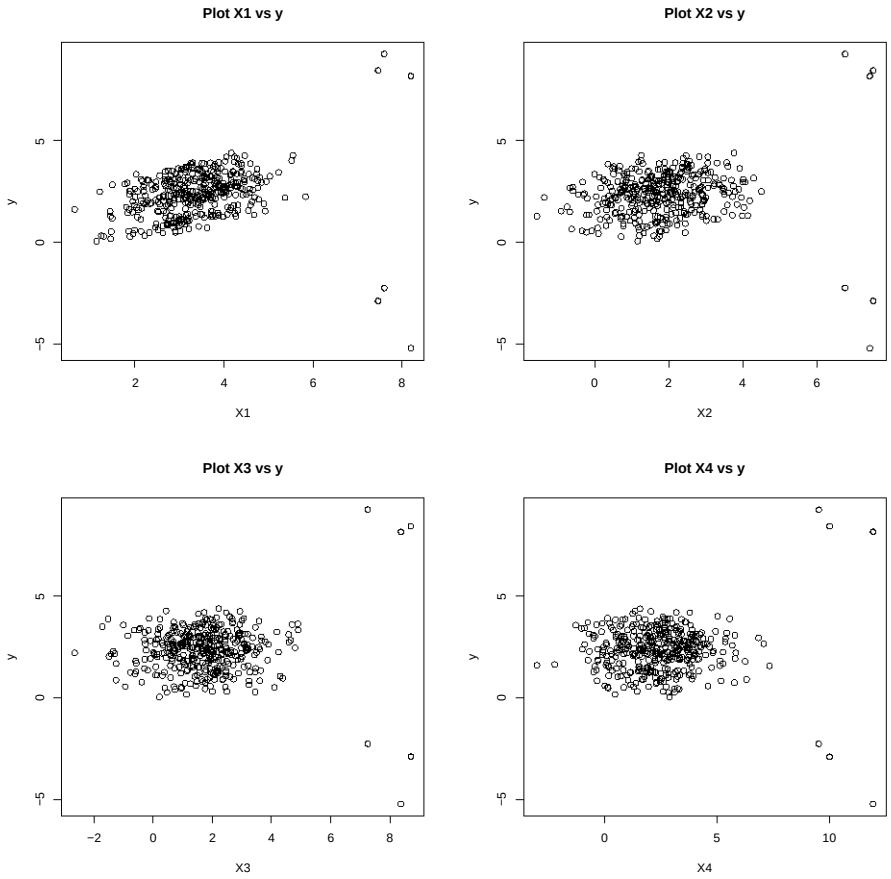


Figure 3.2: Scatterplot of Y versus X_1 (top left), X_2 (top right), X_3 (bottom left) and X_4 (bottom right)

to evaluate the outlier detection performance, and the other two assess robustness properties. The first one is the percentage of the Monte Carlo replications in which all outliers were detected, denoted ALLD. The second criterion is the average over the Monte Carlo simulations of the number of false outliers found by each of these procedures, denoted AFO. In fact, this last criterion attempts to summarize the swamping effect, which occurs when non-outliers are wrongly identified due to the effect of some hidden outliers, see Lawrence [1]. The third criterion is the overall empirical mean squared error (MSE) of the final estimator

$\hat{\gamma}$ obtained by each of the three procedures, defined as

$$\text{MSE}(\hat{\boldsymbol{\gamma}}^{(l)}) = \frac{1}{L}\sum_{l=1}^{L} \|(\hat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma})\|^2, \tag{3.14}$$

where γ is the vector of parameters used to simulate the clean data. Finally, the fourth criterion is the overall empirical median squared error (MNSE), given by

$$\text{MNSE}(\hat{\boldsymbol{\gamma}}^{(l)}) = \text{median}\{\|(\hat{\boldsymbol{\gamma}}^{(l)} - \boldsymbol{\gamma})\|^2, 1 \leq l \leq L\}. \tag{3.15}$$

Hubert and Rousseeuw [20] provided the code for obtaining the RDL_1 estimator and the M-S estimator is implemented in the function $lmRob$ of S-PLUS, version 8.0. Based on the M-S method, Rousseeuw and van Zomeren [44] proposed a rule to classify an observation as a *vertical outlier* or as a *leverage point and vertical outlier*. Plotting standarized residuals (using the normalized MAD) versus robust distances (Mahalanobis distances based on a robust covariance matrix), an observation is regarded as a *vertical outlier* if the absolute value of the standarized residual exceeds 2.5. An observation is a *leverage point and vertical outlier* when it is a *vertical outlier* and at the same time is on the right of the vertical line located at the upper 0.975 percent point of a chi-squared distribution with p degrees of freedom.

Table 3.1 reports the results of the first performance criteria, ALLD, for the three classification rules based on the GPSC, RDL_1 and M-S estimators, under contamination levels of 5%, 10%, 20%, 30% and 40%. Table 3.2 lists the results of the second performance criteria, AFO, for the same three classification rules and contamination levels. Tables 3.3 and 3.4 show the results of the MSE and MNSE respectively for the three estimators and for each percentage of contamination.

Tables 3.1 and 3.2 indicate that for the simulated data, the classifying rule based on the GPSC method achieves a high percentage of correct detection while keeping small the number of observations wrongly identified as outliers (swamping effect). This is true for the two considered contamination scenarios B and C. Furthermore, when the sample is not contaminated by outliers, the GPSC rule presents the lowest AFO as compared with the classifying rules based on the RDL_1 and M-S methods. For contamination scenario B with only vertical outliers, it seems that the RDL_1 and M-S rules wrongly identify as outliers several non-outliers, see Table 3.2. On the other hand, for scenario C, the M-S approach keeps a low AFO for all percentages of contamination.

Concerning now the robustness performance criteria MSE and MNSE, Table 3.3 shows that the GPSC estimator presents better MSE figures than the other two estimators except for the largest percentage of contamination, with the M-S estimator following the GPSC one closely. The MNSE figures of these two estimators are even closer, see Table 3.4.

Simulations were also performed by introducing contamination in several groups of the same size instead of groups of different sizes. Results suggested that the GPSC method works better under contamination B and when this contamination is introduced in groups of medium or large size.

Studies also showed that the GPSC method works better when the groups means can be clearly differentiated, i.e., when the variance of groups effects σ_α^2 is clearly greater than individual error variance σ^2.

Table 3.1: ALLD for the rules based on GPSC, RDL_1 and M-S methods, under contamination scenarios B and C with 5%, 10%, 20%, 30% and 40% of contamination within groups $\mathcal{D}_c = \{1, 5, 7\}$.

	5%		10%		20%		30%		40%	
Method	B	C	B	C	B	C	B	C	B	C
GPSC	100,0	100,0	100,0	100,0	100,0	100,0	99,6	100,0	99,0	100,0
RDL_1	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0
M-S	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Table 3.2: AFO for the rules based on GPSC, RDL_1 and M-S methods, under contamination scenarios A, B and C with 5%, 10%, 20%, 30% and 40% of contamination within groups $\mathcal{D}_c = \{1, 5, 7\}$.

	0%	5%		10%		20%		30%		40%	
Method	A	B	C	B	C	B	C	B	C	B	C
GPSC	1,06	0,99	0,99	0,92	0,90	0,82	0,79	0,71	0,76	0,76	0,72
RDL_1	7,44	6,70	6,69	5,86	5,87	4,51	4,53	3,35	3,36	2,43	2,44
M-S	6,11	5,43	0,17	4,77	0,15	3,65	0,11	2,68	0,09	1,84	0,07

Table 3.3: MSE(%) of the GPSC, RDL_1 and M-S estimators, under contamination scenarios A, B and C with 5%, 10%, 20%, 30% and 40% of contamination within groups $\mathcal{D}_c = \{1, 5, 7\}$.

	0%	5%		10%		20%		30%		40%	
Method	A	B	C	B	C	B	C	B	C	B	C
GPSC	4,79	4,93	4,97	5,06	5,07	5,04	5,05	5,35	5,28	7,36	5,51
RDL_1	9,43	9,67	9,64	9,63	9,60	10,15	9,89	9,74	9,80	10,43	9,91
M-S	5,27	5,32	5,29	5,35	5,23	5,28	5,30	5,47	5,45	5,61	5,60

3.7 Application

From the original AAGIS data set, we consider as outcome the variable $income$, as covariates the variables $hectares$, $crops$, $beef$ and $sheep$ and as grouping vari-

Table 3.4: MNSE(%) of the GPSC, RDL_1 and M-S estimators, under contamination scenarios A, B and C with 5%, 10%, 20%, 30% and 40% of contamination within groups $\mathcal{D}_c = \{1, 5, 7\}$.

	0%	5%		10%		20%		30%		40%	
Method	A	B	C	B	C	B	C	B	C	B	C
GPSC	2,23	2,28	2,22	2,29	2,26	2,26	2,26	2,46	2,50	2,55	2,55
RDL_1	4,10	3,86	4,02	4,47	3,88	4,21	3,89	4,17	4,47	4,12	3,89
M-S	2,42	2,36	2,39	2,28	2,33	2,45	2,47	2,37	2,41	2,67	2,73

able the variable *state*, which gives the state in which the farm is located, with 1 = New South Wales, 2 = Victoria, 3 = Queensland, 4 = South Australia, 5 = Western Australia, 6 = Tasmania, 7 = Northern Territory. If we fit model (3.1) using the raw variables, a histogram of residuals reveals a strongly skewed distribution. Taking logs of the outcome (adding a constant to make it always positive) and the covariates and fitting again the model, a histogram of residuals does not seem far from the normal density but still several outliers appear. Trying to identify the true outliers, we applied the three robust fitting methods considered in this paper. Table 3.5 lists the number of farms remaining in each State after deleting the atypical farms pointed out by the classification rules based on each of the three robust methods. Observe that the rule based on RDL_1 method is the one that eliminates the most quantity of atypical farms over all States, with the largest difference in States 1 and 3. Finally, Table 3.6 reports the final regression parameter estimates provided by each method. Observe that the RDL_1 estimates of the group effects are quite different from the M-S and GPSC counterparts, but the last two are somewhat similar. This might be due to the mentioned swamping effect that could be strongly affecting the RDL_1 estimates. The observed similarity between the M-S and GPSC estimates gives some credibility to these two methods.

Table 3.5: Number of farms remaining in each State after deletion of outliers based on RDL_1, M-S and GPSC methods. M-S^1 refers to *vertical outliers* while M-S^2 refers to *leverage points and vertical outliers*.

State	Original	RDL_1	M-S^1	M-S^2	PSC
1	451	432	436	443	446
2	265	258	257	262	260
3	382	355	358	360	372
4	241	235	234	238	239
5	221	210	210	210	214
6	62	61	60	61	62
7	30	26	28	28	29
Total	1652	1577	1583	1602	1622

Table 3.6: Regression parameter estimates obtained by LS, RDL_1, M-S and GPSC methods.

Parameters	LS	RDL_1	M-S	GPSC
$hectares$	0,335	0,339	0,379	0,374
$crops$	0,169	0,144	0,165	0,164
$beef$	0,079	0,060	0,065	0,066
$sheep$	0,029	0,161	0,022	0,021
$State1$	0,677	0,291	0,588	0,604
$State2$	0,604	0,195	0,490	0,511
$State3$	0,607	0,131	0,523	0,539
$State4$	0,534	0,146	0,426	0,450
$State5$	0,667	0,320	0,582	0,596
$State6$	0,711	0,273	0,633	0,652
$State7$	0,543	0,659	0,363	0,420

3.8 Concluding remarks

This work studies the detection of atypical observations for grouped data following a linear regression model with group effects. We propose to calculate group-

wise principal sensitivity components to detect possibly masked high leverage points (*leverage points*). Then, we fit the model to the remaining data and discard the observations with large residuals (*vertical outliers*). The simulation studies show that our robust procedure presents a high mean percentage of simulations with detection of 100% of true outliers while small number of observations were wrongly regarded as outliers. Particulary, when contamination type B is present, the level of the swamping effect in our robust procedure is the lowest among the three robust methods.

We used the criterion of the minimization of a certain scale of the residuals and then discarded the observations with large standardized residuals according to that scale. However, another alternative would be to approximate the quantiles of the maximum absolute residual by a resampling procedure, then examine each possible candidate and decide whether it is an outlier or not by comparing it with the selected quantile. However, this might be computationally much more intensive.

Chapter 4

Linear model with random effects

4.1 Introduction

This chapter introduces a linear regression model with random group effects, which is a particular case of the linear mixed models that will be introduced in Chapter 6. This model is widely used to analyze clustered data, when the number of clusters is large but there are a small number of observations per cluster. They are are frequently used in many fields such as small area estimation or longitudinal studies because they adequately model the within-cluster correlation (within-subject in longitudinal data) typically present in these type of data. Other fields of application include clinical trials (Vangeneugden et al. [52]) and environmental studies (Wellenius et al. [55]).

Despite the many different applications of these models, still diagnostic methods are not so well developed. Christensen et al. [8] studied case deletion diagnostics. Banerjee and Frees [3] studied case deletion and subject deletion diagnostics. Galpin and Zewotir [15] and [16] extended some diagnostic tools of ordinary linear regression, such as residuals, leverages and outliers to linear mixed models (LMMs) when the variances of the random factors are known. This chapter intro-

duces some of these diagnostics tools.

4.2 Linear model with random effects

Let us consider sample data that come from D different populations groups. Suppose that there are n_d observations from group d, $d = 1, \ldots, D$, where $n = \sum_{d=1}^{D} n_d$ is the total sample size. Denote y_{dj} the value of the study variable for j-th sample unit from d-th group and $\mathbf{x}_{dj}$ a (column) vector containing the values of p auxiliary variables for the same unit. The model at individual level is given by

$$y_{dj} = \mathbf{x}_{dj}^T \boldsymbol{\beta} + u_d + e_{dj} \quad j = 1, \ldots, n_d \quad d = 1, \ldots, D. \tag{4.1}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed parameters, u_d is the random effect of d-th group and e_{dj} is the model error. Random group effects and errors are supposed to be independent with distributions

$$u_d \overset{iid}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2).$$

Observe that under this model, in contrast with model (3.1), the means of the observations are not affected by the group effect u_d since $E(y_{dj}) = \mathbf{x}_{dj}^T \beta$. However, the random group effects induce a (constant) correlation between all pairs of observations in the same group, because $cov(y_{dj}, y_{dk}) = \sigma_u^2$ for $k \neq j$. Still, observations in different groups are uncorrelated. Stacking the elements of the model in columns, we obtain $\mathbf{y} = (y_{11}, y_{12}, \ldots, y_{Dn_D})^T$ of size n, $\mathbf{u} = (u_1, u_2, \ldots, u_D)^T$ of size D and $\mathbf{e} = (e_{11}, e_{12}, \ldots, e_{Dn_D})^T$ of size n. In turn, concatenation of the predictor vectors gives the $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{Dn_D})^T$. Additionally, we define

the $n \times D$ block diagonal matrix

$$\mathbf{Z} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdot & \vdots \\ \vdots & \cdot & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{n_D} \end{pmatrix}$$

where here, $\mathbf{1}_{n_d}$ denotes a vector of ones of size n_d. Then, in matrix notation, the model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_D), \quad \mathbf{e} \sim N(\mathbf{0}, \sigma_e^2 \mathbf{I}_n). \tag{4.2}$$

The expectation and covariance matrix of $\mathbf{y}$ are given by

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad var(\mathbf{y}) = \sigma_u^2 \mathbf{Z}\mathbf{Z}^T + \sigma_e^2 \mathbf{I}_n = \mathbf{V}.$$

which means that

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_u^2 \mathbf{Z}\mathbf{Z}^T + \sigma_e^2 \mathbf{I}_n)$$

Let us define the vector of variance components $\theta = (\sigma_u^2, \sigma_e^2)^T$. When θ is known, Henderson [10] obtained the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ and the Best Linear Unbiased Predictor (BLUP) of $\mathbf{u}$, which are defined respectively as

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, \tag{4.3}$$

$$\tilde{\mathbf{u}} = \sigma_u^2 \mathbf{Z}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \tag{4.4}$$

4.3 Estimation of variance components

The estimator (6.4) and the predictor (6.5) depend on θ, which in practice is unknown and needs to be estimated from sample data. The empirical versions of (6.4) and (6.5), called EBLUE and EBLUP respectively, are obtained by replacing a suitable estimator $\hat{\theta}$ for θ in (6.4) and (6.5) and are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{4.5}$$

$$\hat{\mathbf{u}} = \hat{\sigma}_u^2\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4.6}$$

where the hat over $\mathbf{V}$ indicates that θ has been replaced by its estimator $\hat{\theta}$.

Traditional methods for estimating variance components include those based on the likelihood, namely maximum likelihood (ML) and restricted/residual ML (REML), and a moments method called Henderson method III, see e.g., Searle et al. [47]. However, when outliers are present, these methods may deliver estimators with poor properties. Below we briefly review each of these methods.

4.3.1 Maximum likelihood

Maximum likelihood estimation is usually carried out under the assumption that $\mathbf{y}$ has a multivariate normal distribution. Under this assumption, the joint likelihood is given by

$$f(\boldsymbol{\beta}, \theta|\mathbf{y}) = (2\pi)^{-\frac{n}{2}}|\mathbf{V}|^{-1/2}exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}.$$

The joint log-likelihood is

$$\ell(\boldsymbol{\beta}, \theta|\mathbf{y}) = \ln(f(\boldsymbol{\beta}, \theta|\mathbf{y})) = c - \frac{1}{2}[\ln|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})],$$

where c is denotes a constant. Using the relations

$$\frac{\partial \ln |\mathbf{V}|}{\partial \theta} = tr\left\{\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\right\} \quad \text{and} \quad \frac{\partial \mathbf{V}^{-1}}{\partial \theta} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1},$$

The first order partial derivatives of ℓ with respect to $\boldsymbol{\beta}$, σ_u^2 and σ_e^2 are

$$\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta}, \theta|\mathbf{y})}{\partial \boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial \ell(\boldsymbol{\beta}, \theta|\mathbf{y})}{\partial \sigma_u^2} &= -\frac{1}{2}tr\left\{\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^T\right\} + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\
\frac{\partial \ell(\boldsymbol{\beta}, \theta|\mathbf{y})}{\partial \sigma_e^2} &= -\frac{1}{2}tr\{\mathbf{V}^{-1}\} + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),
\end{aligned}$$

and equating them to zero we obtain the equations

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \mathbf{X}\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \tag{4.7}$$

$$tr\{\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^T\} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{Z}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{4.8}$$

$$tr\{\mathbf{V}^{-1}\} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.9}$$

Solving for $\boldsymbol{\beta}$ in (4.7), we obtain the ML estimating equation for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$

where here $\mathbf{V}$ depends on the ML estimator of $\theta = (\sigma_u^2, \sigma_e^2)^T$. Equations (4.8) and (4.9) do not have analytic solution and need to be solved numerically by iterative methods such as Newton-Raphson or Fisher-scoring.

4.3.2 Restricted maximum likelihood

A criticism of ML estimators of variance components is that they are biased downward, because they do not take into account the loss in degrees of freedom from

the estimation of $\boldsymbol{\beta}$, (Lindstrom and Bates [27]). REML method corrects for this problem by transforming $\mathbf{y}$ into two independent vectors, $\mathbf{y}_1 = K_1\mathbf{y}$ and $\mathbf{y}_2 = K_2\mathbf{y}$. The probability density function of $\mathbf{y}_1$ does not depend on $\boldsymbol{\beta}$ and it holds $E(\mathbf{y}_1) = \mathbf{0}$, which means that $K_1\mathbf{X} = \mathbf{0}$. On the other hand, $\mathbf{y}_2$ is independent of $\mathbf{y}_1$, which means that $K_1\mathbf{V}K_2^T = \mathbf{0}$. The matrix K_1 is chosen to have maximum rank, i.e. $n - p$, so the rank of K_2 is p. The likelihood function of $\mathbf{y}$ is the product of the likelihoods of $\mathbf{y}_1$ and $\mathbf{y}_2$. The variance components coming from the REML approach are the ML estimators of these parameters based on $\mathbf{y}_1$, see [32], [41]. Similarly to the ML case, the obtained equations do not have analytic solutions and need to be solved using iterative techniques such as EM algorithm, Fisher-scoring or Newton-Raphson methods.

Jennrich and Schluchter [22] compared the performances of the three algorithms and noted the following: (1) direct comparison of these algorithms in terms of required computational burden is difficult, because this depend to a large degree of how efficiently the algorithms are coded. (2) Newton-Raphson algorithm, with a quadratic convergence rate, generally converges in a small number of iterations, with a higher cost per iteration. (3) EM method has the lowest cost per iteration, but at times requires a large number of iterations. (4) Fisher-scoring algorithm is intermediate in terms of cost per iteration and required number of iterations. However, its cost per iteration is often not much smaller than that of Newton-Raphson algorithm, whereas Fisher-scoring algorithm sometimes requires a considerably larger number of iterations than Newton-Raphson algorithm. Lindstrom and Bates [27] provided arguments favoring the use of Newton-Raphson method.

4.3.3 Henderson method III

ML and REML estimators of θ are usually based on the assumption that the vector $\mathbf{y}$ has a multivariate normal distribution, although they remain consistent even when normality is not satisfied exactly under some regularity conditions (Jiang, [21]). An alternative method which does not rely on normality and provides explicit formulas for the estimators of the variance components is Henderson method III (H3). This method works as follows. First, consider a linear mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta}$ might contain fixed and random effects. Let us split $\boldsymbol{\beta}$ into two subvectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and define the full model as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}. \tag{4.10}$$

The partition in sum of squares of model (4.10) is given by

$$\begin{aligned}
&\text{SSR}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{y}, \\
&\text{SSE}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{e}^T\mathbf{e} = [(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X})\mathbf{y}]^T[(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X})\mathbf{y}], \\
&\text{SST}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \mathbf{y}^T\mathbf{y},
\end{aligned} \tag{4.11}$$

with their corresponding expected values given by

$$\begin{aligned}
E[\text{SSR}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)] &= tr\left\{ \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix} E(\boldsymbol{\beta}\boldsymbol{\beta}^T) \right\} + \text{rank}(\mathbf{X})\sigma_e^2, \\
E[\text{SSE}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)] &= [n - \text{rank}(\mathbf{X})]\sigma_e^2, \\
E[\text{SST}\ (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)] &= tr\left\{ \begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix} E(\boldsymbol{\beta}\boldsymbol{\beta}^T) \right\} + n\sigma_e^2.
\end{aligned} \tag{4.12}$$

Now consider the reduced model with only $\boldsymbol{\beta}_1$,

$$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon}. \tag{4.13}$$

Analogously, the partition in sum of squares of model (4.13) is given by

$$\begin{aligned}
&\text{SSR }(\boldsymbol{\beta}_1) = \mathbf{y}^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1\mathbf{y},\\
&\text{SSE }(\boldsymbol{\beta}_1) = \epsilon^T\epsilon = [(\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1)\mathbf{y}]^T[(\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1)\mathbf{y}],\\
&\text{SST }(\boldsymbol{\beta}_1) = \mathbf{y}^T\mathbf{y},
\end{aligned} \tag{4.14}$$

with their corresponding expected values

$$\begin{aligned}
E[\text{SSR }(\boldsymbol{\beta}_1)] &= tr\left\{\begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2, \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2 \end{pmatrix} E(\boldsymbol{\beta}\boldsymbol{\beta}^T)\right\} + \text{rank}(\mathbf{X}_1)\sigma_e^2,\\
E[\text{SSE }(\boldsymbol{\beta}_1)] &= tr\{\mathbf{X}^T[\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T]^T[\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T]\mathbf{X}E(\boldsymbol{\beta}\boldsymbol{\beta}^T)\}\\
&\quad + [n - \text{rank}(\mathbf{X})]\sigma_e^2,\\
E[\text{SST }(\boldsymbol{\beta}_1)] &= tr\left\{\begin{pmatrix} \mathbf{X}_1^T\mathbf{X}_1 & \mathbf{X}_1^T\mathbf{X}_2 \\ \mathbf{X}_2^T\mathbf{X}_1 & \mathbf{X}_2^T\mathbf{X}_2 \end{pmatrix} E(\boldsymbol{\beta}\boldsymbol{\beta}^T)\right\} + n\sigma_e^2.
\end{aligned} \tag{4.15}$$

The reduction in sum of squares due to introducing $\mathbf{X}_2$ in the model with only $\mathbf{X}_1$ is

$$\text{SSR}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1) = \text{SSR}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) - \text{SSR}(\boldsymbol{\beta}_1). \tag{4.16}$$

The expectation of this reduction is given by

$$\begin{aligned}
E[\text{SSR }(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)] &= tr\{\mathbf{X}_2^T[\mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T]\mathbf{X}_2E(\boldsymbol{\beta}\boldsymbol{\beta}^T)\}\\
&\quad + [\text{rank}(\mathbf{X}) - \text{rank}(\mathbf{X}_1)]\sigma_e^2.
\end{aligned} \tag{4.17}$$

Now consider model (4.1) and rewrite it as (4.10) taking $\boldsymbol{\beta}_1 = \boldsymbol{\beta}$, $\boldsymbol{\beta}_2 = \mathbf{u}$, $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{X}_2 = \mathbf{Z}$. This method equates the sum of squares $\text{SSR}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ in (4.14) and $\text{SSR}(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1)$ in (4.16) to their expectations in (4.12) and (4.17) respectively, obtaining two equations. Solving for σ_e^2 and σ_u^2 in the resulting equations, we obtain unbiased estimators for σ_e^2 and σ_u^2 (for more details see [47], chapter 5). Let $\hat{\mathbf{e}}$ and $\hat{\boldsymbol{\varepsilon}}$ be the vectors of residuals obtained by fitting the two models (4.10) and (4.13) respectively, considering $\boldsymbol{\beta}_2$ as fixed. If $\text{rank}(\mathbf{X}) = p$ and $\text{rank}(\mathbf{X}|\mathbf{Z}) = p + D$, then the Henderson III estimators of the variance components are given by

$$\hat{\sigma}^2_{e,H3} = \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{e}^2_{dj}}{n-p-D}, \quad \hat{\sigma}^2_{u,H3} = \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{\varepsilon}^2_{dj} - \hat{\sigma}^2_e(n-p)}{tr\left\{\mathbf{Z}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Z}\right\}}, \tag{4.18}$$

where $\hat{e}_{dj}$ is the residual corresponding to observation $(\mathbf{x}_{dj}^T, y_{dj})$ in model (4.10) and $\hat{\epsilon}_{dj}$ is the corresponding in model (4.13).

4.4 Diagnostic methods

Limited work has been done on diagnostic methods for linear mixed models. Christensen et al. [8] considered the case deletion diagnostics and Galpin and Zewotir [16] provided a definition of residuals, leverages and outliers when some variance components are known.

Fitted values of the response variable are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}},$$

and residuals are then

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{R}\mathbf{y},$$

with $\mathbf{R} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$.

Studentized residuals (internal studentization):

$$t_{dj} = \frac{\hat{e}_{dj}}{\sqrt{\text{var}(\hat{e}_{dj})}} = \frac{\hat{e}_{dj}}{\hat{\sigma}_e \sqrt{r_{dj}}}$$

where r_{dj} is the dj-th diagonal element of matrix $\mathbf{R}$ and e_{dj} is the dj-th element of vector $\hat{\mathbf{e}} = \mathbf{R}\mathbf{y}$.

Studentized residuals (external studentization): Let $\hat{\sigma}_{e(dj)}$ denote the estimate of σ_e when the *dj*-th observation is deleted. If $\hat{\sigma}^2_{e(dj)}$ is used in place of $\hat{\sigma}^2_e$ we obtain the *dj*-th externally Studentized residual, given by

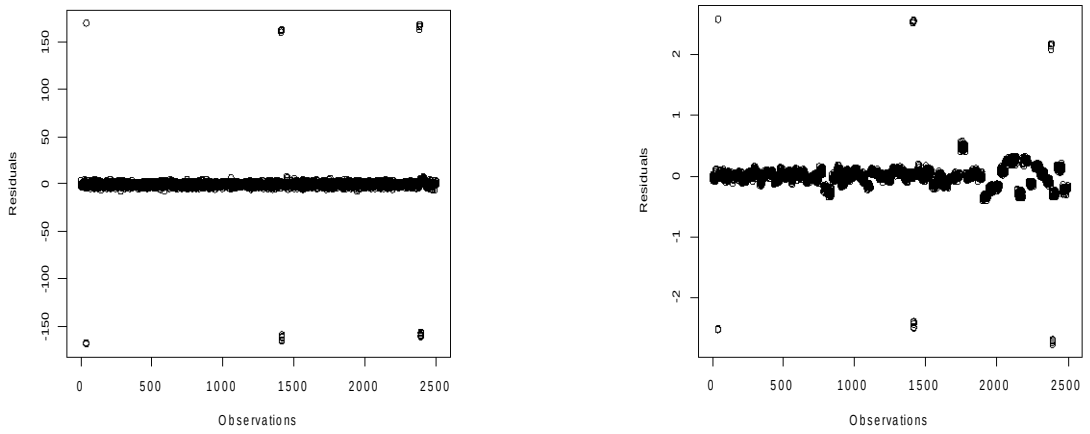
$$t^*_{dj} = \frac{\hat{e}_{dj}}{\hat{\sigma}_{e(dj)} \sqrt{r_{dj}}}.$$

The estimator t^*_{dj} satisfies that $t^{*2}_{dj} \sim \frac{n-1}{n-p-1} F(1, n-p-1)$ where $F(1, n-p-1)$ is an F-distribution with degrees of freedom 1 and $(n-p-1)$.

Note that element the r_{dj} used to standardized residuals depends on the variance components σ^2_e and σ^2_u, which are unknown. When there are outliers, these might affect the estimators of variance components, and these estimators will change the distribution of standarized residuals.

To illustrate this, we have simulated data from model (4.1), with $D = 15$ groups and total sample size $n = 2500$. The theoretical values of the variance components are $\sigma^2_e = 0.5$ and $\sigma^2_u = 0.5$. In order to increase the estimator of the error variance σ^2_e, we introduced atypical data on $\mathbf{y}$ as mean shifts, by increasing the values of the some of the response values by k times the theoretical standard deviation with $k = 5$. Index plots of internally studentized residuals, using the true variance components and the estimated ones, appear in the left and right panels of Figure 4.1 respectively. This example illustrates how the estimation of variance

components affect the studentized residuals. On the right plot obtained with estimated variances, all residuals appear in the interval (-2.5,2.5); as a consequence, using the standard rule applied to these residuals, outlying observations will not be detected.



(a) Variance components known (b) Henderson III

Figure 4.1: Internally studentized residuals (a) using the true variance components and (b) when they are estimated using H3 method.

Leverage effect in the nested-error model

Assuming that θ is known, the vector of predicted values is

$$\tilde{\mathbf{y}} = (\mathbf{I} - \mathbf{R})\mathbf{y} \tag{4.19}$$

This relation evokes the definition of the Hat matrix, as

$$\mathbf{H}_{\tilde{\mathbf{y}}} = \mathbf{I} - \mathbf{R}.$$

The diagonal elements $(1 - r_{dj})$ of this matrix are measures of the leverage effect of the observations and are called *leverages*. Galpin and Zewotir [16] proposed the use of the r_{dj}s to identify influential observations. If r_{dj} approaches zero, this indicates that the corresponding observation has a large leverage effect.

Due to the grouped data structure in linear mixed models with one random factor, it seems more relevant to study the leverage effect of groups instead of that of isolated observations. The leverage effect of group d is defined here as

$$h_d = \overline{\mathbf{x}}_d^T (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \overline{\mathbf{x}}_d, \quad , \quad d = 1, \ldots, D \tag{4.20}$$

where $\overline{\mathbf{x}}_d = n_d^{-1} \sum_{j=1}^{n_d} \mathbf{x}_{dj}$. In practice, $\mathbf{V}$ could be estimated using the robust variance components estimators described in the next chapter.

Chapter 5

Robust fitting of linear models with random effects

5.1 Introduction

This chapter introduces new robust estimators of variance components based on Henderson method III. This method has been chosen for three reasons; first, because it provides explicit formulas for the estimators, avoiding iterative procedures and the need for starting values and reducing the computational time; second, because it does not need any assumption on the shape of the probability of the distribution apart from the existence of first and second order moments; third, the estimation procedure consists simply of solving two standard regression problems. These estimators can later be used to derive robust estimators of regression coefficients. Finally, we describe an application of this procedure to small area estimation, in which the main target is the estimation of the means of areas or domains when the within-area sample sizes are small.

5.2 Robust Henderson method III

Consider the linear regression model with random effects given in (4.1). The estimators of variance components obtained by Henderson method III (H3 estimators) are given by

$$\hat{\sigma}^2_{e,H3} = \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{e}^2_{dj}}{n-(p+D)}, \quad \hat{\sigma}^2_{u,H3} = \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{\varepsilon}^2_{dj} - \hat{\sigma}^2_e(n-p)}{tr\left\{\mathbf{Z}^T[\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Z}\right\}}, \tag{5.1}$$

where $\hat{e}_{dj}$ is the residual corresponding to observation $(\mathbf{x}^T_{dj}, y_{dj})$ in the full model (4.10) with group effects assumed to be fixed and $\hat{\varepsilon}_{dj}$ is the corresponding residual in the reduced model (4.13).

Remark 5.1. Henderson III estimators are scale equivariant, that is,

$$\hat{\sigma}_{e,H3}(cy) = |c|\sigma_{e,H3}(y) \quad \text{and} \quad \hat{\sigma}_{u,H3}(cy) = |c|\sigma_{u,H3}(y).$$

.

Proof. The estimator $\hat{\sigma}^2_{e,H3}$ can be expressed as

$$\hat{\sigma}^2_{e,H3} = \hat{\sigma}^2_{e,H3}(\mathbf{y}) = \frac{SSE(\boldsymbol{\beta}^*)}{n-\text{rank}(\mathbf{X}^*)} = \frac{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}}{n-(p+D)}$$

where $\mathbf{H}^* = \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}$, $\mathbf{X}^* = (\mathbf{X}|\mathbf{Z})$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta}^T, \mathbf{u}^T)^T$.

Then,

$$\begin{aligned}\hat{\sigma}_{e,H3}(c\mathbf{y}) &= \sqrt{\frac{(c\mathbf{y})^T(\mathbf{I}_n-\mathbf{H}^*)(c\mathbf{y})}{n-(p+D)}} \\ &= \sqrt{\frac{c^2\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}}{n-(p+D)}} \\ &= |c|\sqrt{\frac{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}}{n-(p+D)}} \\ &= |c|\hat{\sigma}_{e,H3}(\mathbf{y})\end{aligned}$$

Therefore, the estimator $\hat{\sigma}_{e,H3}$ is scale invariant. Now we check that $\hat{\sigma}_{u,H3}$ is also scale equivariant.

The estimator $\hat{\sigma}^2_{u,H3}$ is given by

$$\hat{\sigma}^2_{u,H3} = \hat{\sigma}^2_{u,H3}(y) = \frac{SSE(\boldsymbol{\beta}) - \hat{\sigma}^2_{e,H3}(n-p)}{tr[\mathbf{Z}^T(\mathbf{I}_n-\mathbf{H})\mathbf{Z}]} = \frac{\mathbf{y}(\mathbf{I}_n-\mathbf{H})\mathbf{y} - \left[\frac{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}}{n-(p+D)}\right](n-p)}{tr[\mathbf{Z}^T(\mathbf{I}_n-\mathbf{H})\mathbf{Z}]}$$

denoting $m = tr[\mathbf{Z}^T(\mathbf{I}_n-\mathbf{H})\mathbf{Z}]$

$$\hat{\sigma}^2_{u,H3} = \frac{1}{m}\left\{\mathbf{y}(\mathbf{I}_n-\mathbf{H})\mathbf{y} - \frac{n-p}{n-(p+D)}\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}\right\}$$

thus,

$$\hat{\sigma}_{u,H3}(\mathbf{y}) = \sqrt{\frac{1}{m}\left\{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H})\mathbf{y} - \frac{n-p}{n-(p+D)}\mathbf{y}^T(I_n-\mathbf{H}^*)\mathbf{y}\right\}}$$

Then,

$$\begin{aligned}
\hat{\sigma}_{u,H3}(c\mathbf{y}) &= \sqrt{\frac{1}{m}\left\{(c\mathbf{y})^T(\mathbf{I}_n-\mathbf{H})(c\mathbf{y})-\frac{n-p}{n-(p+D)}(c\mathbf{y})^T(\mathbf{I}_n-\mathbf{H}^*)(c\mathbf{y})\right\}} \\
&= \sqrt{\frac{c^2}{m}\left\{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H})\mathbf{y}-\frac{n-p}{n-(p+D)}\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}\right\}} \\
&= |c|\sqrt{\frac{1}{m}\left\{\mathbf{y}^T(\mathbf{I}_n-\mathbf{H})\mathbf{y}-\frac{n-p}{n-(p+D)}\mathbf{y}^T(\mathbf{I}_n-\mathbf{H}^*)\mathbf{y}\right\}} \\
&= |c|\hat{\sigma}_{u,H3}(\mathbf{y})
\end{aligned}$$

Therefore, the estimator $\hat{\sigma}_{u,H3}$ is scale invariant.

□

Let us express Henderson III estimators in terms of the means of squared residuals

$$\hat{\sigma}^2_{e,H3}=\frac{n\left[\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{e}^2_{dj}/n\right]}{n-(p+D)},\quad \hat{\sigma}^2_{u,H3}=\frac{n\left[\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{\varepsilon}^2_{dj}/n\right]-\hat{\sigma}^2_e(n-p)}{tr\left\{\mathbf{Z}^T[\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Z}\right\}}, \quad (5.2)$$

We propose to robustify these estimators using, first, robust methods to fit the two models (4.10) and (4.13) and, after that, replacing in (5.2) the means of squared residuals by other robust functions.

Model (4.13) is a standard linear regression model, which can be robustly fitted using any method available in the literature such as L_1 estimation, M estimation or the fast method of Peña and Yohai [34]. Model (4.10) is a model with fixed group effects, which can be robustly fitted using an adaptation of the principal sensibility components method of Peña and Yohai [34] to the grouped data structure. An alternative approach is the M-S estimation of Maronna and Yohai [28].

These fitting methods will provide better residuals $\hat{e}_{dj}$ and $\hat{\varepsilon}_{dj}$, which are in turn used to find robust estimators of the variance components. Below we describe different estimators based on robust functions of these new residuals.

MADH3 estimators: In the two estimators given in (5.2), we substitute the means of squared residuals by the square of the normalized medians of absolute deviations (MAD), given by

$$\text{MAD} = 1.481 \quad \text{median}(|\hat{\xi}_{dj}|, \hat{\xi}_{dj} \neq 0),$$

where $\hat{\xi}_{dj}$ is the residual of observation $(\mathbf{x}_{dj}^T, y_{dj})$ under the corresponding fitted model, either (4.10) or (4.13).

TH3 estimators: Trimming consists of giving zero weight to a percentage of extreme cases. In this case, in the two equations given in (5.2) we trim residuals that are outside the interval (b_1, b_2) with

$$b_1 = q_1 - k(q_3 - q_1) \quad \text{and} \quad b_2 = q_3 + k(q_3 - q_1). \tag{5.3}$$

Here, q_1 and q_2 are the first and third sample quartiles of residuals and k is a constant. Based on results obtained from different simulation studies, we propose to use the constant $k = 2$, just slightly smaller than that one used as outer frontier in the box-plot for detecting outliers.

RH3 estimators: Instead of replacing extreme residuals by zero as in the previous proposal, we can smooth residuals appearing in (5.2) according to an appropriate smoothing function. Here we consider Tukey's biweight function, given by

$$\varphi(x) = x[1 - (x/k)^2]^2, \quad \text{if } |x| \leq k. \tag{5.4}$$

In this case, the robust Henderson III estimators are given by

$$\hat{\sigma}^2_{e,RH3} = \frac{\sigma^2_{e,MAD}\sum_{d=1}^{D}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{dj}/\sigma_{e,MAD})}{n-(p+D)}, \tag{5.5}$$

$$\hat{\sigma}^2_{u,RH3} = \frac{\sigma^2_{u,MAD}\sum_{d=1}^{D}\sum_{j=1}^{n_d}\varphi^2(\hat{\varepsilon}_{dj}/\sigma_{u,MAD}) - \hat{\sigma}^2_{e,RH3}(n-p)}{tr\{\mathbf{Z}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Z}\}}. \tag{5.6}$$

Remark 5.2. The function $h(x) = \sigma_x\varphi(x/\sigma_x)$ is scale invariant, where σ_x is a scale such that $\sigma_{cx} = c\sigma_x, c > 0$. If we consider $\sigma_x = \text{MAD}(x)$, let us verify that

$$\text{MAD}(cx) = c\text{MAD}(x), \quad c > 0.$$

By definition $\text{MAD}(x) = 1.4826\,\text{median}(|x - \text{median}(x)|)$

$$\begin{aligned}
\text{MAD}(cx) &= 1.4826\ \text{median}(|(cx) - \text{median}(cx)|) \\
&= 1.4826\,\text{median}(|c|(x - \text{median}(x))|) \\
&= |c|[1.4826\,\text{median}(|x - \text{median}(x)|)] \\
&= |c|\text{MAD}(x).
\end{aligned}$$

Since $\sigma_{c\,x} = c\,\sigma_x$, we have that

$$h(c\,x) = c\sigma_x\psi\left(\frac{c\,x}{c\sigma_x}\right) = c\sigma_x\psi\left(\frac{x}{\sigma_x}\right) = h(x).$$

Remark 5.3. RH3 estimators of σ^2_e and σ^2_u are scale invariant.

Proof. Consider the estimator $\hat{\sigma}^2_{e,RH3}$

$$\hat{\sigma}^2_{e,RH3} = \frac{\sigma^2_{e,MAD}\sum_{d=1}^{D}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{dj}/\sigma_{e,MAD})}{n-(p+D)} = \sqrt{\frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}h^2(\hat{e_{dj}})}{n-(p+D)}},$$

where $h(\cdot)$ is scale invariant. Therefore, $\hat{\sigma}_{e,RH3}$ is scale invariant.

Let $m = tr\{\mathbf{Z}^T(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Z}\}$. The estimator $\hat{\sigma}^2_{u,RH3}$ is given by

$$\begin{aligned}\hat{\sigma}^2_{u,RH3} &= \frac{1}{m}\left\{\sigma^2_{\varepsilon,MAD}\sum_{d=1}^{D}\sum_{j=1}^{n_d}\varphi^2\left(\frac{\hat{\varepsilon}_{dj}}{\sigma_{\varepsilon,MAD}}\right) - \hat{\sigma}^2_{e,RH3}(n-p)\right\} \\ &= \frac{1}{m}\left\{\sum_{d=1}^{D}\sum_{j=1}^{n_d}h^2(\hat{\varepsilon}_{dj}) - \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}h^2(\hat{e}_{dj})}{n-(p+D)}(n-p)\right\}.\end{aligned}$$

Similarly, since $h(\cdot)$ is scale invariant, $\hat{\sigma}_{u,RH3}$ is scale invariant.

□

5.2.1 Simulation experiment

This section describes a Monte Carlo simulation study that compares the robust estimators of the variance components with the traditional non-robust ones. For this, we generated data coming from $D = 10$ groups. The group sample sizes n_d, $d = 1, \ldots, D$ were respectively 20, 20, 30, 30, 40, 40, 50, 50, 60 and 60, with a total sample size of $n = 400$. We considered $p = 4$ auxiliary variables, and they were generated from normal distributions with means and standard deviations coming from a real data set from the Australian Agricultural and Grazing Industries Survey. Thus, the values of the four auxiliary variables were generated respectively as $X_1 \sim N(3.3, 0.6)$, $X_2 \sim N(1.7, 1.2)$, $X_3 \sim N(1.7, 1.6)$ and $X_4 \sim N(2.4, 2.6)$. The simulation study is based on $L = 500$ Monte Carlo replicates. In each iteration, we generated group effects as $u_d \overset{iid}{\sim} N(0, \sigma^2_u)$ with $\sigma^2_u = 0.25$. Similarly, we generated errors as $e_{dj} \overset{iid}{\sim} N(0, \sigma^2_e)$ with $\sigma^2_e = 0.25$. Then we generated the model responses y_{dj}, $j = 1, \ldots, n_d$, $d = 1, \ldots, D$, from model (4.1). Observe that in principle there is no contamination. Finally, we introduced contamination according to three different scenarios:

A. *No contamination.*

B. *Groups with a mean shift:* A subset $\mathcal{D}_c \subseteq \{1, 2, \ldots, D\}$ of groups was selected for contamination. For each selected group $d \in \mathcal{D}_c$, half of the observations were replaced by $c_{d1} = \bar{y}_d + k\, s_{Y,d}$ and the other half by $c_{d2} = \bar{y}_d - k\, s_{Y,d}$ with $k = 5$, where $\bar{y}_d$ and $s_{Y,d}$ are respectively the mean and the standard deviation of the outcome for the clean data in d-th group. This increases the between group variability σ_u^2.

C. *Groups with high variability:* A small percentage of contaminated observations was introduced in each selected group $d \in \mathcal{D}_c$, similarly as described in Scenario B. This increases the within group variability σ_e^2.

With each Monte Carlo sample, we fitted the two models (4.10) and (4.13) using respectively the GPSC method described in Chapter 3 and the robust procedure of Peña and Yohai [34]. Then, we calculated the traditional estimators H3, ML and REML, and the proposed robust estimators, MADH3, TH3 and RH3. After the $L = 500$ replicates, we computed the empirical bias and mean squared error (MSE) of the estimators.

Table 5.1 reports the resulting empirical bias and percent MSE of each estimator under Scenario A, without contamination. Observe in that table that in absence of outlying observations, the traditional non-robust estimators, H3, ML and REML, provide the minimum MSE, but the robust alternatives TH3 and RH3 are not too far away from them. However, under Scenario B with full groups contaminated with a mean shift (Tables 5.2 and 5.3), the estimators ML, REML and H3 of σ_u^2 increase considerably their MSE. The estimator TH3 achieves the minimum MSE, followed by RH3. Under Scenario C with contamination introduced to make the within cluster variability increase (Tables 5.4 and 5.5), now the estimators ML, REML and H3 of σ_e^2 increase considerably their MSE whereas the robust estimators resist quite well.

Table 5.1: Theoretical values $\sigma_u^2 = \sigma_e^2 = 0.25$. Scenario 0: No contamination.

Method	Estimators		Bias		MSE $\times 10^2$	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$
H3	0,24	0,25	-0,0081	0,0014	1,43	0,03
ML	0,22	0,25	-0,0298	-0,0011	1,16	0,03
REML	0,25	0,25	-0,0046	0,0014	1,32	0,03
MADH3	0,25	0,25	0,0041	0,0018	2,33	0,09
TH3	0,23	0,25	-0,0189	-0,0019	1,04	0,04
RH3	0,24	0,23	-0,0136	-0,0179	1,25	0,06

Table 5.2: Theoretical values $\sigma_u^2 = \sigma_e^2 = 0.25$. Scenario B: One outlying group.

Method	Estimators		Bias		MSE $\times 10^2$	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$
H3	1,28	0,24	1,0286	-0,0095	123,73	0,04
ML	1,15	0,24	0,9000	-0,0120	123,27	0,04
REML	1,28	0,24	1,0285	-0,0096	123,38	0,04
MADH3	0,44	0,23	0,1884	-0,0169	7,84	0,10
TH3	0,24	0,24	-0,0089	-0,0142	1,25	0,05
RH3	0,46	0,22	0,2106	-0,0277	6,04	0,10

Table 5.3: Theoretical values $\sigma_u^2 = \sigma_e^2 = 0.25$. Scenario B: Two outlying groups.

Method	Estimators		Bias		MSE $\times 10^2$	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$
H3	2,79	0,23	2,5375	-0,0242	715,98	0,08
ML	2,13	0,22	1,8807	-0,0266	495,49	0,10
REML	2,37	0,23	2,1179	-0,0242	500,14	0,08
MADH3	1,10	0,21	0,8529	-0,0437	91,67	0,25
TH3	0,27	0,22	0,0227	-0,0319	2,13	0,13
RH3	0,76	0,21	0,5088	-0,0412	31,52	0,19

Table 5.4: Theoretical values $\sigma_u^2 = \sigma_e^2 = 0.25$. Scenario C: 10% of atypical observations shared among groups.

Method	Estimators		Bias		MSE $\times 10^2$	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$
H3	0,23	0,60	-0,0175	0,3512	1,47	12,58
ML	0,21	0,60	-0,0397	0,3450	1,23	12,15
REML	0,24	0,60	-0,0144	0,3512	1,35	12,58
MADH3	0,28	0,27	0,0253	0,0198	2,78	0,14
TH3	0,24	0,25	-0,0073	-0,0012	1,17	0,04
RH3	0,22	0,30	-0,0266	0,0487	1,22	0,26

Table 5.5: Theoretical values $\sigma_u^2 = \sigma_e^2 = 0.25$. Scenario C: 20% of atypical observations shared among groups

Method	Estimators		Bias		MSE $\times 10^2$	
	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$	$\hat{\sigma}_u^2$	$\hat{\sigma}_e^2$
H3	0,22	0,93	-0,0268	0,6814	1,50	47,19
ML	0,20	0,92	-0,0489	0,6719	1,32	45,89
REML	0,23	0,93	-0,0236	0,6814	1,39	47,19
MADH3	0,30	0,29	0,0473	0,0406	3,48	0,29
TH3	0,25	0,25	0,0045	0,0003	1,27	0,04
RH3	0,21	0,37	-0,0400	0,1151	1,18	1,35

5.2.2 Conclusions

This work introduces three robust versions of H3 estimators called MADH3, TH3 and RH3 estimators. These estimators are obtained by first, fitting in a robust way the two submodels (4.10) and (4.13) and, then, replacing the means of squared residuals in H3 estimators by other robust functions of the residuals coming from those robust fittings. In simulations we have analyzed the robustness of our proposed estimators under two different contamination scenarios: when the between groups variability is increased by including a mean shift in some of the groups, and when the within group variability is increased by introducing given percentages of outliers within the groups. The new robust estimator RH3 achieves great efficiency under both types of contamination and at the same time preserves good efficiency when there is not contamination.

5.3 Robust estimation of regression coefficients

This section deals with robust estimation of regression coefficients using the estimators of variance components introduced above. These estimators are then used to derive robust predictors of the means in small areas.

5.3.1 Small area estimators

Small area estimation is usually done under the setup of finite population. Thus, we have a population U of size N that is assumed to be partitioned into D subpopulations $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$ called small areas. Particular quantities of interest are the means of the small areas,

$$\overline{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}, \quad d = 1 \ldots, D$$

A sample s_d of size n_d is drawn from U_d, $d = 1, \ldots, D$. We assume that the model

holds for all population units, that is, for units in the sample and out of the sample. Under this setup, the target area means are random. Therefore, is it common to say predicting $\overline{Y}_d$ rather than estimating $\overline{Y}_d$. The mean of small area d can be split into two terms, one for the sample elements an the other for the out-of-sample elements, obtaining a linear combination of the sample mean $\overline{y}_{s_d}$ and the out-of-sample mean $\overline{y}_{s_d^c}$.

$$\overline{Y}_d = \frac{1}{N_d}\left[\sum_{j\in s_d} y_{dj} + \sum_{j\in s_d^c} y_{dj}\right] = \frac{n_d}{N_d}\overline{y}_{s_d} + \left(\frac{n_d}{N_d}\overline{y}_{s_d^c}\right), \quad d = 1\ldots, D$$

When studying outliers in finite population inference, the existing literature is developed exclusively under one of the following assumptions:

Assumption 1. Non representative outliers: We assume that atypical observations appear only in the sample but not in the non-sample part of the population. Then, it seems natural to project the working model into the entire non-sampled part of the population. Chambers [42] call these type of outliers non-representative outliers. In this case, the appropriate methods for estimating model parameters are called *Robust Projective*, meaning that they project sample non-outlier behavior on to the non-sampled part of the population.

Assumption 2. Representative outliers We assume that atypical observations appear in the sample and non-sample part of the population. In this case, robust projective methods will provide biased estimators of the small area means; therefore, it is necessary to correct for this bias using an appropriate correction factor.

Next section introduces two robust projective methods given in the literature, Fellner's approach and Sinha and Rao's procedure.

5.3.2 Previous robust procedures

Fellner's approach

Fellner [56] derived robust estimators of variance components and regression coefficients $\boldsymbol{\beta}$, together with a robust predictor of $\mathbf{u}$, which could in turn be used to derive a robust EBLUP.

The joint probability density function of $\mathbf{y}$ is given by

$$f(\boldsymbol{\beta}, \theta|\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{5.7}$$

Similarly, the joint density function of $\mathbf{u} = (u_1, \ldots, u_D)^T$ is

$$g(\mathbf{u}; \sigma_u^2) = (2\pi\sigma_u^2)^{-D/2} \exp\{-\mathbf{u}^T\mathbf{u}/2\}.$$

Assuming θ known, the BLUE of $\boldsymbol{\beta}$ and the BLUP of $\mathbf{u}$ can be obtained simultaneously by maximizing the joint loglikelihood of $\mathbf{y}$ and $\mathbf{u}$, $\ln f(\boldsymbol{\beta}, \theta|\mathbf{y}, \mathbf{u}) = \ln f(\theta|\mathbf{y}) + \ln g(\mathbf{u})$, with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$. The resulting system of normal equations is given by

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X}/\sigma_e^2 & \mathbf{X}^T\mathbf{Z}/\sigma_e^2 \\ \mathbf{Z}^T\mathbf{X}/\sigma_e^2 & \mathbf{I}/\sigma_u^2 + \mathbf{Z}^T\mathbf{Z}/\sigma_e^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y}/\sigma_e^2 \\ \mathbf{Z}^T\mathbf{y}/\sigma_e^2 + (\mathbf{I}/\sigma_u^2)\mathbf{0}_D \end{bmatrix}$$

Fellner's method is based in the idea of replacing in these equations, observations y_{di} and random effects u_d that are far from their predicted values $\hat{y}_{di} = \mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}} + \hat{u}_d$ and $\hat{u}_d$ by what he called pseudo-observations. More explicitly, Fellner's method solves the system

$$\begin{bmatrix} \mathbf{X}^T\mathbf{X}/\sigma_e^2 & \mathbf{X}^T\mathbf{Z}/\sigma_e^2 \\ \mathbf{Z}^T\mathbf{X}/\sigma_e^2 & I/\sigma_u^2 + \mathbf{Z}^T\mathbf{Z}/\sigma_e^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\mathbf{y}^*/\sigma_e^2 \\ \mathbf{Z}^T\mathbf{y}^*/\sigma_e^2 + (\mathbf{I}/\sigma_u^2)\mathbf{0}_D^* \end{bmatrix}, \tag{5.8}$$

where $\mathbf{y}^* = (y^*_{di}, i = 1, \ldots, n_d, d = 1, \ldots, D)$ with $y^*_{di} = \mathbf{x}^T_{dj}\hat{\boldsymbol{\beta}} + \hat{u}_d + \sigma_e\psi(\hat{e}_{dj}/\sigma_e)$ and $\mathbf{0}^*_D = (\hat{u}_d - \sigma_u\phi(\hat{u}_d/\sigma_u); d = 1, \ldots, D)$ and ψ es an odd, monotonic and bounded function such as Huber's psi function.

Equations (5.8) assume that variance components are known, but Fellner [56] also gave REML equations for variance components which, solved jointly with (5.8), yield also a robust estimator of $\boldsymbol{\beta}$ together with a robust predictor of $\mathbf{u}$. For this, he proposes to robustify REML equations in the form

$$\begin{aligned}
\hat{\sigma}^2_u &= \{h(D - v^*)\}^{-1}\hat{\sigma}_u \sum_{d=1}^{D} \psi^2(\hat{u}_d/\hat{\sigma}_u), \\
\hat{\sigma}^2_e &= \{h(n - p - D + v^*)\}^{-1}\hat{\sigma}_e \sum_{d=1}^{D} \psi^2(\hat{e}_{dj}/\hat{\sigma}_e),
\end{aligned}$$

where h is an appropriately chosen constant to adjust for the bias in $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_e$ at the normal distribution. This leads to $h = E\{\psi^2(X)\}$, where $X \sim N(0, 1)$.

REBLUP estimators

Sinha and Rao [49] proposed a two-step procedure for constructing robust estimators of model parameters. The steps of the procedure are the following:

- **Step 1**. The estimators $\hat{\boldsymbol{\beta}}^{SR}$ and $\hat{\theta}^{SR}$ are obtained simultaneously based on robustified ML equations.

- **Step 2**. The predictor $\hat{\mathbf{u}}^{SR}$ is obtained using the estimators of Step 1.

In Step 1, the ML equations for $\boldsymbol{\beta}$ and θ are defined by

$$\begin{aligned} \mathbf{X}^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) &= \mathbf{0}, \\ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_\ell}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - tr\left\{\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_\ell}\right\} &= \mathbf{0}, \quad \ell = 1,2, \end{aligned}$$

where θ_ℓ is the ℓ-th element of $\theta = (\sigma_u^2, \sigma_e^2)^T$.

If some fitted values $\hat{y}_{dj} = \mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}}$ are unusually different from the corresponding observed values y_{dj}, then we have the indication of apparent outliers in the data. To handle outliers in the response values, they proposed robustified ML equations in the form

$$\begin{aligned} \mathbf{X}^T\mathbf{V}^{-1}U^{\frac{1}{2}}\Psi(\mathbf{r}) &= \mathbf{0}, \\ \Psi(\mathbf{r})^TU^{\frac{1}{2}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_\ell}\mathbf{V}^{-1}U^{\frac{1}{2}}\Psi(\mathbf{r}) - \text{tr}\left\{K\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_\ell}\right\} &= \mathbf{0}, \quad \ell = 1,2, \end{aligned}$$

where

$\mathbf{r} = U^{-\frac{1}{2}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, $U = diag(\mathbf{V})$, $K = E\{\psi_b^2(X)\}\mathbf{I}_n$ with $X \sim N(0,1)$, $\Psi(u) = (\psi_b(u_1), \psi_b(u_2), \ldots)^T$ with $\psi_b(u) = u \cdot \min(1, \frac{b}{|u|})$ and $b = 1.345$.

The complete algorithm for robust estimation of $\boldsymbol{\beta}$ and θ is:

(i) Choose starting values $\boldsymbol{\beta}^{(0)}$ and $\theta^{(0)}$. Set $m = 0$.

(ii) (a) Calculate $\boldsymbol{\beta}^{(m+1)}$. (b) Calculate $\theta^{(m+1)}$. (c) Set $m = m + 1$.

(iii) Repeat until convergence is achieved. Denote the estimates at convergence as $\hat{\boldsymbol{\beta}}^{SR}$ and $\hat{\theta}^{SR}$.

In Step 2, the predictor $\hat{\mathbf{u}}^{SR}$ is obtained using the estimators of $\boldsymbol{\beta}$ and θ obtained in Step 1 and solving the following robustified equation

$$\hat{\sigma}_e \, \mathbf{Z}^T \Psi \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})/\hat{\sigma}_e \right\} - \hat{\sigma}_u \Psi(\mathbf{u}/\hat{\sigma}_u) = 0$$

Sinha and Rao [49] proposed to solve this equation using the Newton-Raphson method. Finally, the Robust EBLUPs (REBLUPs) of the small area means are given by

$$\hat{\overline{Y}}_d^{SR} = \frac{1}{N_d} \left(\sum_{j \in s_d} y_{dj} + \sum_{j \epsilon s_d^c} \hat{y}_{dj}^{SR} \right), \quad d = 1, \ldots, D$$

where $\hat{y}_{dj}^{SR} = \mathbf{x}_{dj}^T \hat{\boldsymbol{\beta}}^{SR} + \hat{u}_d^{SR}$.

Some comments

The Newton-Raphson procedure is a commonly used iterative method for the solution of nonlinear equations. To solve the equation $h(t) = 0$, at each iteration the function h is linearized in the sense that it is replaced by its Taylor expansion of order one about the current approximation. Let us denote by t_m the m-th approximation. Then the next value is the solution of

$$h(t^m) + h'(t^m)(t^{m+1} - t^m) = 0$$

that is,

$$t^{m+1} = t^m - \frac{h(t^m)}{h'(t^m)}$$

If the procedure converges, the convergence is very fast; but it is not guaranteed to converge. If h' is not bounded away from zero, the denominator may become very small, making the sequence t^m unstable unless the initial value t^0 is very near to the solution (Maronna et al., [29]).

5.3.3 Procedure using RH3

We propose a two-step procedure that provides robust estimators of model parameters based on the robust estimators of variance components given in (5.2).

- **Step 1**. Obtain the estimator $\hat{\theta}^{RH3}$ using the robustified version of Henderson Method III given in (5.5) and (5.6).
- **Step 2**. Obtain the estimator $\hat{\boldsymbol{\beta}}^{RH3}$ and the predictor $\hat{\mathbf{u}}^{RH3}$ similarly as in Sinha and Rao [49], solving the robustified normal equations (5.8).

Then, the new robust EBLUPs, called here RH3-EBLUPs of the small area means are given by

$$\hat{\overline{Y}}_d^{RH3} = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\epsilon s_d^c} \hat{y}_{dj}^{RH3}\right), \quad d = 1, \ldots, D$$

where $\hat{y}_{dj}^{RH3} = \mathbf{x}_{dj}^T\hat{\boldsymbol{\beta}}^{RH3} + \hat{u}_d^{RH3}$.

5.3.4 Simulation experiment

In this simulation study we generated data coming from $D = 30$ groups. Concerning the group sample sizes, half of them were taken of size $n_d = 10$ and the other half of size $n_d = 20$, with a total sample size of $n = 450$. We considered $p = 4$ auxiliary variables, and they were generated from normal distributions with means and standard deviations coming from a real data set from the Australian Agricultural and Grazing Industries Survey. More concretely, the values of the four auxiliary variables were generated respectively as $X_1 \sim N(3.31, 0.68)$, $X_2 \sim N(1.74, 1.23)$, $X_3 \sim N(1.70, 1.65)$ and $X_4 \sim N(2.41, 2.61)$.

The number of Monte Carlo samples was $L = 200$. In each replicate, group effects were generated as $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ with $\sigma_u^2 = 1$. Similarly, individual errors were generated as $e_{dj} \overset{iid}{\sim} N(0, \sigma_e^2)$ with $\sigma_e^2 = 1$. Finally, model responses y_{dj}, $j = 1, \ldots, n_d$, $d = 1, \ldots, D$, were generated from model (4.1). Using each Monte Carlo sample, the two models (4.10) and (4.13) were fitted robustly using respectively the M-S estimator of Maronna and Yohai [28] and the PSC method of Peña and Yohai [34]. We assume that outliers are representative and use the correction factor proposed by Joingo et al. [Joingo D]. Firstly, data are generated without contamination. After that, contamination is introduced according to the following scenarios:

- **Type 0**. No contamination
- **Type 1**. Outlying areas: For each selected outlying domain, we substitute all their sample observations y_{dj} by the constant $C_1 = \bar{Y}_d + c \cdot \sqrt{\frac{\sum_{j=1}^{N_d}(y_{dj} - \bar{Y}_d)^2}{N_d}}$, where $c = 4$ and $\bar{Y}_d = \frac{1}{N_d}\sum_{j=1}^{N_d} y_{dj}$.
- **Type 2**. Outlying individuals within areas: We replace some observations within selected domains by C_1 and some others by $C_2 = \bar{Y}_d - c \cdot \sqrt{\frac{\sum_{j=1}^{N_d}(y_{dj} - \bar{Y}_d)^2}{N_d}}$.

To compare several predictors of the prediction of the small area means, we use the following measures averaged over areas

Average Absolute Relative Bias ($\overline{ARB}$):

$$\overline{ARB} = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{L}\sum_{t=1}^{L}\left(\frac{\hat{\bar{Y}}_d - \bar{Y}_d}{\bar{Y}_d}\right)\right|$$

Average Relative Root MSE ($\overline{RRMSE}$):

$$\overline{RRMSE} = \frac{1}{D}\sum_{d=1}^{D}\frac{\overline{MSE}(\hat{\bar{Y}}_d)^{\frac{1}{2}}}{\bar{Y}_d}$$

Method	Bias		MSE	
	σ_u^2	σ_e^2	σ_u^2	σ_e^2
ML	-0,044	0,070	0,160	0,125
REML	-0,125	0,141	0,247	0,195
RH3	-0,174	0,075	0,279	0,142

Table 5.6: Scenario Type 0: No contamination

Parameter	ML		REML		RH3	
	Bias	MSE	Bias	MSE	Bias	MSE
β_0	-0,037	0,264	-0,033	0,312	-0,034	0,321
β_1	0,316	0,014	0,314	0,015	0,312	0,014
β_2	0,001	0,012	0,001	0,013	0,003	0,013
β_3	-0,007	0,004	-0,006	0,005	-0,008	0,005

Table 5.7: Scenario Type 0: No contamination

5.3.5 Conclusions

This work compares two ways to estimate regression coefficients in the linear with random effects. Then, these estimators were used to derive robust predictors of the means in small areas. Our simulation studies show that the new robust procedure RH3 gets the best results in the case of outlying areas at the same time good efficiency when there is not contamination.

Method	$\overline{ARB}$	$\overline{RRMSE}$
EBLUP	0,3667	0,3825
REBLUP	0,4015	0,5056
RH3-EBLUP	0,3843	0,4884

Table 5.8: Scenario Type 0: No contamination

Method	Bias		MSE	
	σ_u^2	σ_e^2	σ_u^2	σ_e^2
ML	2,346	-0,022	6,248	0,119
REML	0,838	0,335	1,430	0,362
RH3	0,437	-0,167	0,586	0,227

Table 5.9: Scenario Type 1: One outlying domain.

Parameter	ML		REML		RH3	
	Bias	MSE	Bias	MSE	Bias	MSE
β_0	0,250	0,319	0,092	0,308	0,087	0,306
β_1	0,318	0,015	0,324	0,016	0,324	0,016
β_2	-0,013	0,012	-0,005	0,013	-0,006	0,013
β_3	-0,003	0,004	-0,007	0,005	-0,008	0,005

Table 5.10: Scenario Type 1: One outlying domain.

Method	$\overline{ARB}$	$\overline{RRMSE}$
EBLUP	0,4161	0,5301
REBLUP	0,4192	0,5251
RH3-EBLUP	0,4193	0,5248

Table 5.11: Scenario Type 1: One outlying domain.

Method	Bias		MSE	
	σ_u^2	σ_e^2	σ_u^2	σ_e^2
ML	5,027	-0,267	26,706	0,186
REML	3,205	0,478	15,848	0,541
RH3	2,386	-0,319	6,076	0,277

Table 5.12: Scenario Type 1: Two outlying domains.

Parameter	ML		REML		RH3	
	Bias	MSE	Bias	MSE	Bias	MSE
β_0	0,637	0,688	0,336	0,453	0,307	0,459
β_1	0,304	0,015	0,317	0,018	0,318	0,018
β_2	-0,016	0,013	-0,009	0,014	-0,008	0,015
β_3	-0,009	0,004	-0,012	0,005	-0,010	0,005

Table 5.13: Scenario Type 1: Two outlying domains.

Method	$\overline{ARB}$	$\overline{RRMSE}$
EBLUP	0,4162	0,6296
REBLUP	0,4316	0,5652
RH3-EBLUP	0,4338	0,5502

Table 5.14: Scenario Type 1: Two outlying domains.

Method	Bias		MSE	
	σ_u^2	σ_e^2	σ_u^2	σ_e^2
ML	-0,095	0,959	0,173	1,046
REML	-0,159	0,363	0,296	0,349
RH3	-0,216	0,364	0,293	0,322

Table 5.15: Scenario Type 2: 10% outlying observations within groups.

Parameter	ML		REML		RH3	
	Bias	MSE	Bias	MSE	Bias	MSE
β_0	-0,014	0,342	-0,028	0,337	-0,018	0,323
β_1	0,316	0,016	0,315	0,016	0,314	0,015
β_2	0,001	0,009	0,002	0,014	0,004	0,009
β_3	-0,006	0,005	-0,006	0,005	-0,008	0,005

Table 5.16: Scenario Type 2: 10% outlying observations within groups.

Method	$\overline{ARB}$	$\overline{RRMSE}$
EBLUP	0,4417	0,5286
REBLUP	0,3963	0,5002
RH3-EBLUP	0,3849	0,4881

Table 5.17: Scenario Type 2: 10% outlying observations within groups.

Method	Bias		MSE	
	σ_u^2	σ_e^2	σ_u^2	σ_e^2
ML	-0,180	1,912	0,184	3,783
REML	-0,214	0,604	0,293	0,567
RH3	-0,232	0,575	0,286	0,554

Table 5.18: Scenario Type 2: 20% outlying observations within groups.

Parameter	ML		REML		RH3	
	Bias	MSE	Bias	MSE	Bias	MSE
β_0	0,005	0,367	-0,028	0,352	-0,018	0,353
β_1	0,306	0,020	0,316	0,018	0,314	0,017
β_2	-0,006	0,015	-0,002	0,015	-0,001	0,015
β_3	-0,007	0,006	-0,008	0,005	-0,009	0,005

Table 5.19: Scenario Type 2: 20% outlying observations within groups.

Method	$\overline{ARB}$	$\overline{RRMSE}$
EBLUP	0,4265	0,5440
REBLUP	0,3895	0,4920
RH3-EBLUP	0,3825	0,4845

Table 5.20: Scenario Type 2: 20% outlying observations within groups.

Chapter 6

Robust fitting of linear mixed models

6.1 Introduction

Chapter 4 of this dissertation studied linear models with random effects, which are a particular case of linear mixed models in which only one random factor or source of variation (apart from individual error) is considered in the model. These models are used for clustered or longitudinal data. However, sometimes data show a more complex structure such as clustering at different levels or cross-classification. Moreover, we might consider in the model other sources of variation such as variation in time and/or space. Linear mixed models are used when data present multiple sources of variation. They are used in many different fields of application such as Biology, Econometrics and Engineering and have received considerable attention both from a practical and theoretical point of view, see e.g. McCulloch and Searle [31], Verbeke and Molenberghs [53], [43], SRS Rao [50], Muller and Stewart [38], Sahai and Ojeda [46], Rao JNK [23] and Demidenko [13]. Part of their success can be due to the fact that these models avoid problems of multidimensionality, because only few parameters need to be estimated, in contrast with fixed effects models in which a large number of regression parameters must be estimated. Since regression coefficients are deemed as random variables,

these models can be seen as a compromise between the frequentist and Bayesian approaches.

6.2 Linear mixed model

Consider that the vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ of observations from our study variable obeys the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \cdots + \mathbf{Z}_r\mathbf{u}_r + \mathbf{e}\,, \tag{6.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients for the explanatory variables and $\mathbf{u}_i = (u_{i1}, \ldots, u_{iD_i})^T$ is the vector containing the effects of the D_i levels of the i-th random factor, $i = 1, \ldots, r$. These random factors are variables that affect the variability of our data. For simplicity of language, the vector $\mathbf{u}_i$ itself will be called i-th random factor. The vector $\mathbf{e} = (e_1, \ldots, e_n)^T$ contains the individual errors and $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ and $\mathbf{X}$ are design matrices of orders $n \times D_1, \ldots, n \times D_r$ and $n \times p$ respectively. Matrix $\mathbf{Z}_i$ contains only zeros and ones, with only one 1 in each row and at least one 1 in each column, $i = 1, \ldots, r$. All random components in the model, $\mathbf{u}_1, \ldots, \mathbf{u}_r$ and $\mathbf{e}$ are independent and they are usually assumed to satisfy

$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma_e^2\mathbf{I}_n), \quad \mathbf{u}_i \sim N_{D_i}(\mathbf{0}, \sigma_{u_i}^2\mathbf{I}_{D_i}),\ i = 1, \ldots, r.$$

Estimability of the model parameters requires usual assumptions, namely that the number of observations is larger than the number of parameters $n \geq p + r + 1$, that there are not multicollinearity problems in the columns of $\mathbf{X}$, that is, $\text{rank}(\mathbf{X}) = p$, that the columns of $\mathbf{X}$ are not collinear with the effects of the random factors, that is, $\text{rank}(\mathbf{X}|\mathbf{Z}_i) > p$, $i = 1, \ldots, r$, and finally that the effects of one of the random factors is not confounded with the effects of the other factors,

that is, $\mathbf{Z}_i\mathbf{Z}_i^T$ and $\mathbf{I}$ are linearly independent,

$$\alpha_0\mathbf{I} + \sum_{i=1}^{r} \alpha_i \mathbf{Z}_i\mathbf{Z}_i^T = 0 \Longrightarrow \alpha_i = 0,\ i = 0, 1, \ldots, r.$$

From model assumptions, it holds that

$$\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),\ \text{with } \mathbf{V} = \sigma_e^2\mathbf{I}_n + \sum_{i=1}^{r} \sigma_{u_i}^2 \mathbf{Z}_i\mathbf{Z}_i^T.$$

Let us define the matrix $\mathbf{Z} = (\mathbf{Z}_1|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)$ and the vector $\mathbf{u} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_r^T)^T$. Then, the model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{6.2}$$

which fits the notation used in Chapter 4. Defining additionally the vector of variance components $\theta = (\sigma_e^2, \sigma_{u_1}^2, \ldots, \sigma_{u_r}^2)^T$, the likelihood is given by

$$f(\theta|\mathbf{y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{6.3}$$

As in Chapter 4, the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$ and the Best Linear Unbiased Predictor (BLUP) of $\mathbf{u}$ obtained by Henderson [10] are given by

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{6.4}$$

$$\tilde{\mathbf{u}} = \sigma_u^2\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \tag{6.5}$$

but they depend on the vector of variance components θ, which is unknown and needs to be estimated.

6.3 Henderson method III

Consider the linear mixed model defined above,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}_1. \tag{6.6}$$

Model (6.6) will be called **full model**. Now consider the following r reduced models (there would be different sets of reduced models from which to estimate the variance components, for further details see Searle et al., [47]).

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \sum_{i=2}^{r} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}_2, \\
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \sum_{i=3}^{r} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}_3, \\
&\vdots \\
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e}_{r+1}.
\end{aligned} \tag{6.7}$$

Consider the sum of squared residuals from the full model, the reduction in regression sum of squares due to introducing $\mathbf{u}_1$ in a model with $\mathbf{u}_2, \ldots, \mathbf{u}_r$, the same when introducing $\mathbf{u}_1$ and $\mathbf{u}_2$ in a model with $\mathbf{u}_3, \ldots, \mathbf{u}_r$, etc, and the same when introducing $\mathbf{u}_1, \ldots, \mathbf{u}_r$ in a model with $\boldsymbol{\beta}$, that is

$$\begin{aligned}
&\text{SSE}(\boldsymbol{\beta}, \mathbf{u}_1, \ldots, \mathbf{u}_r), \\
&\text{SSR}(\mathbf{u}_1 | \boldsymbol{\beta}, \mathbf{u}_2, \ldots, \mathbf{u}_r) = \text{SSR}(\boldsymbol{\beta}, \mathbf{u}_1, \ldots, \mathbf{u}_r) - \text{SSR}(\boldsymbol{\beta}, \mathbf{u}_2, \ldots, \mathbf{u}_r), \\
&\text{SSR}(\mathbf{u}_1, \mathbf{u}_2 | \boldsymbol{\beta}, \mathbf{u}_3 \ldots, \mathbf{u}_r) = \text{SSR}(\boldsymbol{\beta}, \mathbf{u}_1, \ldots, \mathbf{u}_r) - \text{SSR}(\boldsymbol{\beta}, \mathbf{u}_3, \ldots, \mathbf{u}_r), \\
&\quad \vdots \\
&\text{SSR}(\mathbf{u}_1, \ldots, \mathbf{u}_r | \boldsymbol{\beta}) = \text{SSR}(\boldsymbol{\beta}, \mathbf{u}_1, \ldots, \mathbf{u}_r) - \text{SSR}(\boldsymbol{\beta}),
\end{aligned} \tag{6.8}$$

Taking expectation on each of the equations in (6.8), we obtain

$$\begin{aligned}
E[\text{SSE}(\boldsymbol{\beta}, \mathbf{u}_1, \ldots, \mathbf{u}_r)] &= [n - \text{rank}(\mathbf{X}|\mathbf{Z}_1|\ldots|\mathbf{Z}_r)]\sigma_e^2, \\
E[\text{SSR}(\mathbf{u}_1|\boldsymbol{\beta}, \mathbf{u}_2, \ldots, \mathbf{u}_r)] &= tr\{\mathbf{Z}_1^T\mathbf{M}_1\mathbf{Z}_1\}\sigma_{u_1}^2 + [\text{rank}(\mathbf{X}|\mathbf{Z}_1|\ldots|\mathbf{Z}_r) \\
&\quad - \text{rank}(\mathbf{X}|\mathbf{Z}_2|\ldots|\mathbf{Z}_r)]\sigma_e^2, \\
E[\text{SSR}(\mathbf{u}_1, \mathbf{u}_2|\boldsymbol{\beta}, \mathbf{u}_3, \ldots, \mathbf{u}_r)] &= tr\{\mathbf{Z}_1^T\mathbf{M}_2\mathbf{Z}_1\}\sigma_{u_1}^2 + tr\{\mathbf{Z}_2^T\mathbf{M}_2\mathbf{Z}_2\}\sigma_{u_2}^2 \\
&\quad + [\text{rank}(\mathbf{X}|\mathbf{Z}_1|\ldots|\mathbf{Z}_r) - \text{rank}(\mathbf{X}|\mathbf{Z}_3|\ldots|\mathbf{Z}_r)]\sigma_e^2, \\
&\vdots \\
E[\text{SSR}(\mathbf{u}_1, \ldots, \mathbf{u}_r|\boldsymbol{\beta})] &= \sum_{i=1}^{r} tr\{\mathbf{Z}_i^T\mathbf{M}_r\mathbf{Z}_i\}\sigma_{u_i}^2 + [\text{rank}(\mathbf{X}|\mathbf{Z}_1|\ldots|\mathbf{Z}_r) - \text{rank}(\mathbf{X})]\sigma_e^2,
\end{aligned} \tag{6.9}$$

where

$$\begin{aligned}
\mathbf{M}_1 &= \mathbf{I}_n - (\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)[(\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)^T(\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)]^{-1}(\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)^T, \\
\mathbf{M}_2 &= \mathbf{I}_n - (\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)[(\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)^T(\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)]^{-1}(\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)^T, \\
\mathbf{M}_3 &= \mathbf{I}_n - (\mathbf{X}|\mathbf{Z}_4|\cdots|\mathbf{Z}_r)[(\mathbf{X}|\mathbf{Z}_4|\cdots|\mathbf{Z}_r)^T(\mathbf{X}|\mathbf{Z}_4|\cdots|\mathbf{Z}_r)]^{-1}(\mathbf{X}|\mathbf{Z}_4|\cdots|\mathbf{Z}_r)^T, \\
&\vdots \\
\mathbf{M}_r &= \mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T.
\end{aligned}$$

Equating the expectations in (6.9) to the corresponding sums of squares and solving for $\sigma_e^2, \sigma_{u_1}^2, \ldots, \sigma_{u_r}^2$ in the resulting equations, we obtain the Henderson III

estimators of the variance components, given by

$$
\begin{aligned}
\hat{\sigma}^2_{e,H3} &= \frac{\sum_{d=1}^{D}\sum_{j=1}^{n_d}\hat{e}^2_{1,dj}}{n-\text{rank}(\mathbf{X}|\mathbf{Z}_1|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)}, \\
\hat{\sigma}^2_{u_1,H3} &= \frac{\sum_{d=1}^{D_1}\sum_{j=1}^{n_d}\hat{e}^2_{2,dj}-[n-\text{rank}(\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)]\hat{\sigma}^2_e}{\text{tr}\{\mathbf{Z}_1^T\mathbf{M}_1\mathbf{Z}_1\}}, \\
\hat{\sigma}^2_{u_2,H3} &= \frac{\sum_{d=1}^{D_2}\sum_{j=1}^{n_d}\hat{e}^2_{3,dj}-[n-\text{rank}(\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)]\hat{\sigma}^2_e-\text{tr}\{\mathbf{Z}_1^T\mathbf{M}_2\mathbf{Z}_1\}\hat{\sigma}^2_{u_1}}{\text{tr}\{\mathbf{Z}_2^T\mathbf{M}_2\mathbf{Z}_2\}}, \\
&\vdots \\
\hat{\sigma}^2_{u_r,H3} &= \frac{\sum_{d=1}^{D_r}\sum_{j=1}^{n_d}\hat{e}^2_{r+1,dj}-[n-\text{rank}(\mathbf{X})]\hat{\sigma}^2_e-\sum_{i=1}^{r-1}\text{tr}\{\mathbf{Z}_i^T\mathbf{M}_r\mathbf{Z}_i\}\hat{\sigma}^2_{u_i}}{\text{tr}\{\mathbf{Z}_r^T\mathbf{M}_r\mathbf{Z}_r\}},
\end{aligned}
\tag{6.10}
$$

where $\hat{e}_{1,dj}$ is the residual corresponding to observation $(\mathbf{x}_{dj}, y_{dj})$, obtained by fitting model (6.6) but regarding all factors $\mathbf{u}_i$ as fixed and $\hat{e}_{2,dj}, \ldots, \hat{e}_{r+1,dj}$ are the analogous residuals obtained by fitting the reduced models (6.7) respectively, with all factors regarded as fixed.

6.4 Robust Henderson method III

This section provides an extension of the robust Henderson method III introduced in Section 4.3.3 to linear mixed models with several random factors.

Applying a similar approach as in Section 4.3.3, the robust Henderson III esti-

mators of the variance components $\hat{\sigma}^2_{u_1}, \ldots, \hat{\sigma}^2_{u_r}$ and σ^2_e are given by

$$
\begin{aligned}
\hat{\sigma}^2_{e,RH3} &= \frac{\sigma^2_{e,MAD}\sum_{d=1}^{D}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{1,dj}/\sigma_{e,MAD})}{n - \text{rank}(\mathbf{X}|\mathbf{Z}_1|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)}, \\
\hat{\sigma}^2_{u_1,RH3} &= \frac{\sigma^2_{e_1,MAD}\sum_{d=1}^{D_1}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{2,dj}/\sigma_{e_1,MAD}) - [n - \text{rank}(\mathbf{X}|\mathbf{Z}_2|\cdots|\mathbf{Z}_r)]\hat{\sigma}^2_{e,RH3}}{\text{tr}\{\mathbf{Z}_1^T\mathbf{M}_1\mathbf{Z}_1\}}, \\
\hat{\sigma}^2_{u_2,RH3} &= \frac{\sigma^2_{e_2,MAD}\sum_{d=1}^{D_2}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{3,dj}/\sigma_{e_2,MAD}) - [n - \text{rank}(\mathbf{X}|\mathbf{Z}_3|\cdots|\mathbf{Z}_r)]\hat{\sigma}^2_{e,RH3} - \text{tr}\{\mathbf{Z}_1^T\mathbf{M}_2\mathbf{Z}_1\}\hat{\sigma}^2_{u_1,RH3}}{\text{tr}\{\mathbf{Z}_2^T\mathbf{M}_2\mathbf{Z}_2\}}, \\
&\vdots \\
\hat{\sigma}^2_{u_r,RH3} &= \frac{\sigma^2_{e_r,MAD}\sum_{d=1}^{D_r}\sum_{j=1}^{n_d}\varphi^2(\hat{e}_{r+1,dj}/\sigma_{e_r,MAD}) - [n - \text{rank}(\mathbf{X})]\hat{\sigma}^2_{e,RH3} - \sum_{i=1}^{r-1}\text{tr}\{\mathbf{Z}_i^T\mathbf{M}_r\mathbf{Z}_i\}\hat{\sigma}^2_{u_i,RH3}}{\text{tr}\{\mathbf{Z}_r^T\mathbf{M}_r\mathbf{Z}_r\}},
\end{aligned}
\tag{6.11}
$$

where $\varphi(x)$ is Tukey's biweight function and $\sigma_{e,MAD}, \sigma_{e_1,MAD}, \ldots, \sigma_{e_r,MAD}$ are the median of absolute deviations of residuals obtained by fitting the full and each of the reduced models respectively.

Bibliography

[1] AJ, L. (1995). Deletion influence and masking in regression. *Journal of the Royal Statistical Society B*, 57:181–189.

[2] Andrews DF, P. D. (1978). Finding the outliers that matter. *Journal of the Royal Statistical Society. Series B*, 40:85:93.

[3] Banerjee M, F. E. (1997). Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association*, 92:999–1005.

[4] Belsley DA, Kuh E, W. R. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. *Wiley*.

[5] Chambers RL, Pratesi M, S. N. T. N. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17:393–411.

[6] Chambers RL, T. N. (2006). M-quantile models for small area estimation. *Biometrika*, 93:255–268.

[7] Chatterjee S, H. A. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Scrence*, 1:379–416.

[8] Christensen R, Pearson LM, J. W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34:38–45.

[9] Cook RD, W. S. (1982). Residuals and influence in regression. *Chapman and Hall*.

[10] CR, H. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

[11] DL, D. (1982). Breakdown propierties of multivariate location estimators. *Ph.D. qualifying paper, Harvard University*.

[12] Donoho DL, H. P. (1983). The notation of breakdown-point. *A Festschrift Erich L. Lehmann*, page 157–184.

[13] E, D. (2004). Mixed models. theory and applications. *Wiley*.

[14] FR, H. (1971). A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 6:1887–1896.

[15] Galpin JS, Z. T. (2005). Influence diagnostics for linear mixed models. *Journal of Data Science*, 3:153–177.

[16] Galpin JS, Z. T. (2007). A unified approach on residuals, leverages and outliers in the linear mixed models. *Test*, 16:58–75.

[17] Hampel F, Ronchetti E, R. P. S. W. (1986). *Robust statistics: The approach based on influence functions*, volume 1. Wiley, New York.

[18] Hoaglin DC, W. R. (1978). The hat matrix in regresssion and anova. *Journal of the American Statistical Association*, 1:17–22.

[19] Hubert M, R. P. (1996). Robust regression with a categorical covariable. *Robust Statistics, Data Analysis, and Computer Intensive Methods, Lecture Notes in Statistics*, 109:215–224.

[20] Hubert M, R. P. (1997). Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference*, 57:153–163.

[21] J, J. (1996). Reml estimation: asymptotic behavior and related topics. *The Annals of Statistics*, 24:255–286.

[22] Jennrich, RI, S. M. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42:805–820.

[23] JNK, R. (2003). Small area estimation. *Wiley*.

[Joingo D] Joingo D, Haziza D, D. P. Controlling the bias of robust small area estimators. *Working paper, Université of Montréal*.

[25] Kianifard F, S. W. (1989). Using recursive residuals, calculated on adaptively ordered observations, to identify outliers in linear regression. *Biometrics*, 45:571–585.

[26] Kianifard F, S. W. (1990). A monte carlo comparison of five procedures for identifying outliers in linear regression. *Communications in Statistics, Theory and Methods*, 19:1913–1938.

[27] Lindstrom MJ, B. D. (1988). Newtom-rhapson and em algorithm for linear mixed-effects model for repeated-measures. *Journal of the American Statistical Association*, 83:1014–1022.

[28] Maronna, R. and Yohai, V. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89:197–214.

[29] Maronna R, Martin D, Y. V. (2006). Robust statistics. theory and methods. *Wiley*.

[30] Maronna RA, Y. V. (2000). Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89:197–214.

[31] McCulloch Ch, S. S. (2001). Generalized, linear and mixed models. *Wiley*.

[32] Patterson HD, T. R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

[33] Peña D, Y. V. (1995). The detection of influential subsets in linear regression by using a influece matrix. *Journal of the Royal Statistical Society. Series B*, 57:145–156.

[34] Peña D, Y. V. (1999). A fast procedure for outlier diagnostics in large regression problems. *Journal of the American Statistical Association. Theory and Methods*, 94:434–445.

[35] PJ, H. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:73–101.

[36] PJ, H. (1981). Robust statistics. *Wiley, New York*.

[37] PJ, R. (1985). Multivariate estimation with high breakdown point. *In W. Grossmann Pflug G, Vincze T, Wertz W. Eds. Mathematical Statistics and Applications. B Reidel, Dordrecht. The Netherlands*, pages 283–297.

[38] PW, M. K. S. (2006). Linear models theory univariate, multivariate and mixed models. *Willey*.

[39] RD, C. (1977). Detection of influential observations in linear regression. *Technometrics*, 19:15–18.

[40] RE, W. (1982). Influence functions and regression diagnostics. *Modern Data Analysis*, page 149–169.

[41] Richardson AM, W. A. (1995). Robust restricted maximum likelihood in mixed linear models. *Biometrics*, 51:1429–1439.

[42] RL, C. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81:1063–1069.

[43] RM, H. (1993). A robust approach to the analysisi of repeated measures. *Biometrics*, 49:715–720.

[44] Rousseeuw PJ, v. Z. B. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85:633–639.

[45] Rousseeuw PJ, Y. V. (1984). Robust regression by means of s-estimators. *Robust and Nonlinear Time Series. Lectures Notes in Statistics*, 26:256–272.

[46] Sahai H, O. M. (2003). Analysis of variance for random models. *Birkhaüser*.

[47] Searle SR, Casella G, M. C. (1992). Variance components. *Wiley*.

[48] Simonoff JS, H. A. (1993). Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88:1264–1272.

[49] Sinha SK, R. J. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37:381–399.

[50] SRS, P. (1997). Variance components estimation mixed models methodologies and applications. *Chapman and Hall*.

[51] TA, W. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54:427–450.

[52] Vangeneugden T, Laenen A, G. H. R. D. M. G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, 25:13–30.

[53] Verbeke G, M. G. (2009). Linear mixed models for longitudinal data. *Springer*.

[54] VJ, Y. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656.

[55] Wellenius GA, Yeh GY, C. B. S. H. P. R. M. M. (2007). Effects of ambient air pollution on functional status in patients with chronic congestive heart failure: a repeated-measures study. *Environmental Health*, pages 6–26.

[56] WH, F. (1986). Robust estimation of variance components. *Technometrics*, 28:51–60.